

Chapter 4

SARS non-structural protein 12 and ESPRIT

In this chapter the experimental procedure for a directed evolution technique known as Expression of Soluble Protein by Random Incremental Truncation (ESPRIT) is described. The applicability of the ESPRIT method was tested on a biomedically important coronavirus target for which knowledge-based construct design approaches have given no useful expression conditions: the RNA-dependent RNA polymerase (NSP-12) from SARS coronavirus (SARS-CoV). The aim of the experiment was to ‘trap’ domains of the SARS-CoV NSP-12 protein that express solubly in *E. coli* and are thereby suitable for crystallisation and structure determination by X-ray crystallography. Part of this work was carried out at the European Molecular Biology Laboratory (EMBL) in Grenoble using the equipment and protocols of Darren Hart’s group.

4.1. Generation of the ESPRIT libraries

4.1.1. Vector preparation

A codon-optimised SARS-CoV NSP-12 gene (2.796 kb, obtained from GeneART, Regensburg/Germany) was PCR-amplified with primers containing appropriate restriction sites (upstream *AscI*; downstream *NotI*). The resultant PCR product and the screening vector pHAR-1119 were cut with restriction enzymes *AscI* and *NotI* (New England Biolabs) and ligated. The assembled vector was grown and purified from *E. coli* and verified by sequencing. At each end of the inserted NSP-12 gene, there are two main restriction sites: on the vector side, sites *NsiI* or *AatII*, that leave a 3' overhang that is ExoIII resistant; on the

target end side, sites *NotI* or *AscI* that leave a 5' overhang that is an ExoIII substrate (**Fig. 4.1**). In addition, cutting the vector with *AatII* and *AscI* (so as to digest the NSP-12 gene from the 5'-end with ExoIII) removes a stop codon between NSP-12 and a His₆-tag sequence provided by the vector. Upon religation 1 in 3 clones will be in frame with this tag. Similarly, cutting the vector with *NotI* and *NsiI* (so as to digest the NSP-12 gene from the 3'-end with ExoIII) removes a stop codon between NSP-12 and a biotin acceptor peptide (BAP tag, see Chapter 1, section 1.2.1 and **Fig. 1.15**) sequence provided by the vector. Again, 1 in 3 clones will be in frame with the BAP tag upon religation. Therefore, in a bi-directionally truncated library only 1 in 9 clones will be in frame with both tags after religation.

Figure 4.1. Restriction enzyme sites of NSP-12 gene inserted in pHAR1119 vector.

The pHAR1119 vector provides an upstream T7 promoter sequence (yellow) and His₆-TEV tag (orange, cleavable with TEV protease) and a downstream biotin acceptor peptide (BAP) sequence (blue). The vector also provides restriction sites *AatII* and *AscI* after the His₆-TEV tag, and *NotI* and *NsiI* before the BAP tag. Between these sites are stop codon sequences (red) that are removed upon digestion with pairs of restriction enzymes. * indicates a restriction site that will leave a 5' overhang that is sensitive to exonuclease III (ExoIII, pink). ExoIII-sensitive sites ensure digestion into the NSP-12 gene rather than into the vector when cleaving with *AatII/AscI* to digest from the 5'-end and *NotI/NsiI* to digest from 3'-end.



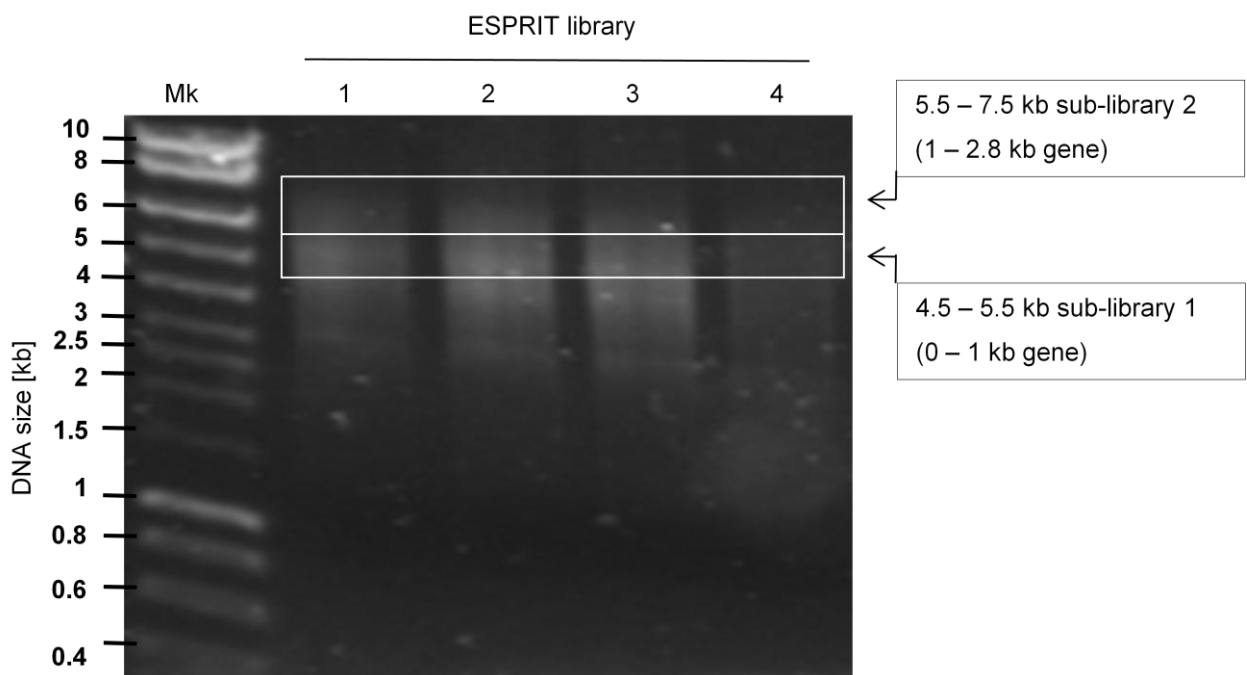
4.1.2. Bi-directional library generation

To create a bi-directionally truncated library, the gene was first truncated from its 5'-end. 10µg of vector DNA was extracted and purified by phenol-chloroform extraction (Sambrook *et al.*, 1989). The vector was sequentially digested using *Aat*II and *Asc*I restriction enzymes (New England Biolabs), and the DNA concentration measured. 4µg of doubly-cut vector (in a volume of 120 µL) was then digested with *Exo*III (New England Biolabs) for 1 h at 22°C. During the course of the reaction, samples were removed at regular intervals (1µL/min) and quenched in 3M NaCl at 4°C. The sample was further digested with Mung Bean Nuclease (New England Biolabs), polished with *Pfu* polymerase (New England Biolabs) and run on a 0.5% agarose gel. The smear on the gel spanning from vector plus full-length insert (7.357 kb) to vector minus full-length insert (4.561 kb) was excised from the gel, purified using the QIAEX II kit (QIAGEN), and ligated using the Rapid DNA ligation kit (Roche Diagnostics) following the manufacturer's instructions for both kits.

This 5'-truncated library was then subjected to a second round of truncation, this time from the 3'-end of the NSP-12 gene. To do this 10µg of the isolated vector DNA library was sequentially digested with *Nsi*I and *Not*I restriction enzymes (New England Biolabs), rather than with *Aat*II and *Asc*I restriction enzymes (New England Biolabs) used to generate the 5'-truncations. Digestion with *Nsi*I and *Not*I opens the vector at the 3'-end of the gene where it can then be truncated from the *Not*I site by *Exo*III. The protocol is essentially the same as for the 5'-truncation however when excising DNA from the 0.5% agarose gel the vector fragments were cut into two sub-libraries (**Fig. 4.2**): 4.561 kb – 5.5 kb fragments (0–1 kb gene truncations) and 5.5 kb – 7.357 kb fragments (1–2.796 kb gene truncations). By size-fractionating it is possible to scan the NSP-12 gene for specific domain sizes of interest.

Figure 4.2. Size-fractionation of ESPRIT library.

0.5% agarose gel of the resulting bi-directional library from ExoIII digestion. The library was loaded across four lanes (1-4) and a smear of DNA of different size ranges can be seen. Mk is the Hyperladder I (Bioline) DNA ladder and white boxes indicate where sub-libraries 1 and 2 were cut out of the gel. The agarose gel was made in 1x TAE buffer and supplemented with a 1 in 10,000 dilution of 10,000X SYBR Safe DNA gel stain (Invitrogen) which allowed visualisation of the bands with a blue light transilluminator.



The DNA in each sub-library was purified using the QIAEX II kit (QIAGEN) and ligated using the Rapid DNA ligation kit (Roche Diagnostics) as described previously. The ligation mixture was then transformed into OMNImaxII cells (Invitrogen) by heat-shock at 42°C for 30 s, then returned to ice for 2 min before recovering in 1 mL Super Optimal Broth (SOC) medium for 1 h at 37°C (no antibiotic selection). 10 and 100 µL aliquots were then plated out on Petri dishes containing LB-Agar supplemented with kanamycin (50 µg/mL). These plates were incubated overnight at 37°C and the remaining SOC cell mix was placed at 4°C.

To test the success of the library generation, 24 colonies from each sub-library were picked and single colony-PCR was carried out using T7 forward and T7 reverse primers which bind to regions of the screening vector flanking the NSP-12 gene (**Fig. 4.3, A and B**). The same 24 colonies were also sequenced (GATC-Biotech) to assess the coverage of constructs across the length of the NSP-12 gene (**Fig. 4.3, C**). The colony forming units per millilitre (CFU/mL) was calculated from the plates so as to determine the volume of SOC cell mix (stored at 4°C) to plate out on much larger LB-Agar Qtrays (Genetix) supplemented with kanamycin (50 µg/mL) to yield greater than 30,000 colonies. These colonies were then harvested and the vector DNA isolated using the Midiprep kit (QIAGEN) following the manufacturer's instructions.

Figure 4.3. (overleaf). Quality testing by colony PCR and sequencing.

A. Colony PCR for sub-library 1. The 24 colonies screened show an acceptable size range from ~200 bp to ~1 kb on the 1.5% agarose gel. **B.** Colony PCR for sub-library 2. The 24 colonies screened show an acceptable size range from ~1 kb to ~3 kb (full length) on the 1% agarose gel. Hyperladder I (Bioline) was used as a marker (Mk) to assess the size range. For A and B, the agarose gels were made in 1x TAE buffer and supplemented with a 1 in 10,000 dilution of 10,000X SYBR Safe DNA gel stain (Invitrogen) which allowed visualisation of the DNA bands with a blue light transilluminator. **C.** Sequencing results. The same 24 clones from colony PCR for each sub-library were sequenced (GATC-Biotech) to assess the size distribution and spread across the gene. The distribution is shown along the length of the 932αα SARS NSP-12 protein: sub-library 1 (green); sub-library 2 (blue).

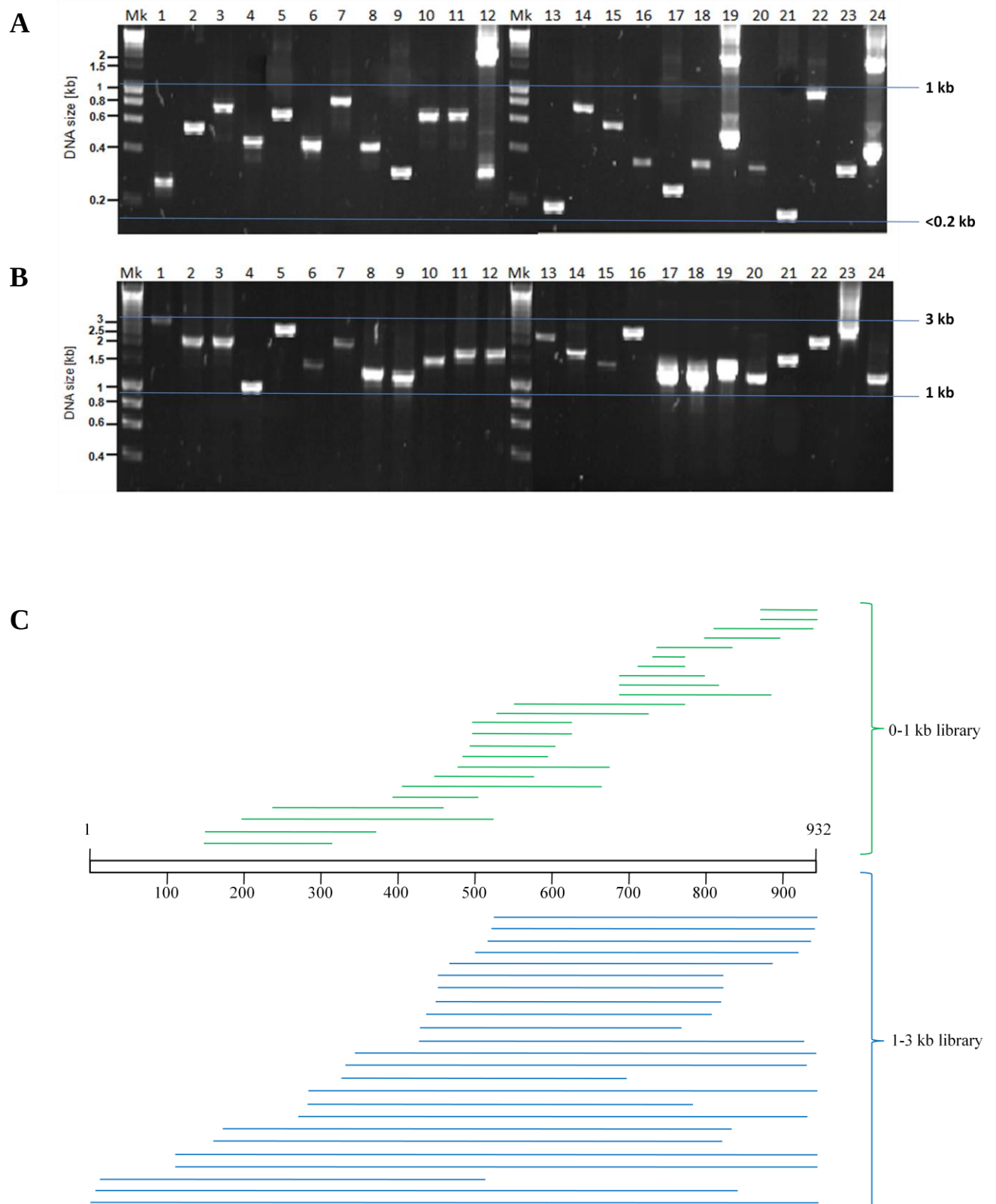


Figure 4.3.

4.2. Library screening – EMBL (Darren Hart Laboratory)

Using the quality-tested library, *E. coli* BL21 RIL (DE3) cells were transformed and plated out on large LB-Agar Qtrays (Genetix) supplemented with 50 µg/mL kanamycin. Individual colonies were picked into LB-medium in 72 384-well plates, using a picker-gridder robot (KBiosystems); in total, 27,648 colonies were picked for sub-library 2 (1 kb to full-length NSP-12 gene fragments) and grown overnight at 37°C in a HiGro high-throughput plate incubator (GeneMachines, Inc.). Too few colonies were obtained from sub-library 1 (smaller NSP-12 gene truncations) and further work on this sub-library was discontinued at this stage. Glycerol stocks of the overnight cultures of colonies from sub-library 2 were prepared by replica-plating and frozen at -80°C.

The picked colonies were gridded into a colony array on a nitrocellulose filter membrane (Amersham Biosciences) on LB-Agar supplemented with 50 µg/mL kanamycin and 50 mM biotin, and incubated at 37°C overnight. For induction, the filter membrane was transferred to an LB-Agar plate supplemented with 50 µg/mL kanamycin, 50 mM biotin and 1 mM IPTG and incubated for 4-6 h at 25°C. The colony array was then lysed on the filter membrane using published protocols (Bussow *et al.*, 1998) and finally washed with PBS supplemented with 0.05% Tween-20 (PBS-T) prior to blocking overnight at 4°C with Superblock blocking buffer in PBS-T (Thermo Scientific). The nitrocellulose membrane was then hybridised with two fluorescent conjugates; all antibodies were made up in 10-fold diluted Superblock buffer. His-tagged protein was detected using mouse anti-hexahistidine (anti-His₆) antibody as the primary antibody and goat anti-mouse antibody conjugated to a fluorescent dye (Alexa Fluor 594; Molecular Probes, Cat. No. A-11032) as the secondary. Biotinylated protein was detected using streptavidin conjugated to a different fluorescent dye (Alexa Fluor 488; Molecular Probes, Cat. No. A-32361). Two fluorescent scans on the same membrane were

taken and were digitally recorded using a Typhoon fluorescent imager (Amersham Biosciences). Alexa-Fluor 594 was excited with green lasers (532 nm) and emission visualised with 610 BP 30 filters that transmit light between 596-625 nm (Alexa-Fluor 594 emission max. = 617 nm; red); Alexa-Fluor 488 was excited with blue laser (488 nm) and emission visualised with 526 SP filters that transmit light below 526 nm (Alexa-Fluor 488 emission max. = 519 nm; green). An example result is shown in **Fig. 4.4**.

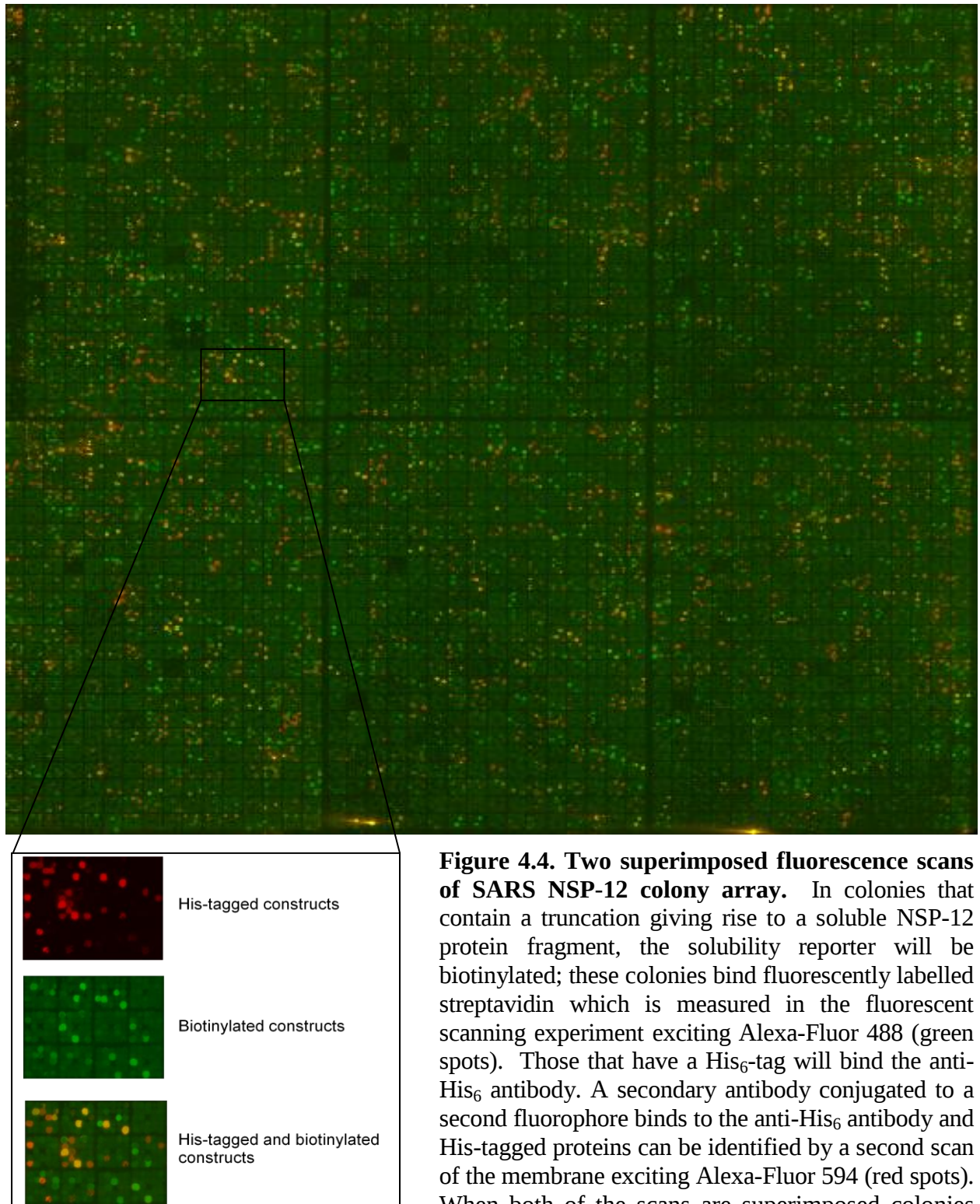


Figure 4.4. Two superimposed fluorescence scans of SARS NSP-12 colony array. In colonies that contain a truncation giving rise to a soluble NSP-12 protein fragment, the solubility reporter will be biotinylated; these colonies bind fluorescently labelled streptavidin which is measured in the fluorescent scanning experiment exciting Alexa-Fluor 488 (green spots). Those that have a His₆-tag will bind the anti-His₆ antibody. A secondary antibody conjugated to a second fluorophore binds to the anti-His₆ antibody and His-tagged proteins can be identified by a second scan of the membrane exciting Alexa-Fluor 594 (red spots). When both of the scans are superimposed colonies containing gene truncations that give rise to a soluble biotinylated protein that can be affinity purified by the presence of a His₆-tag show up in yellow.

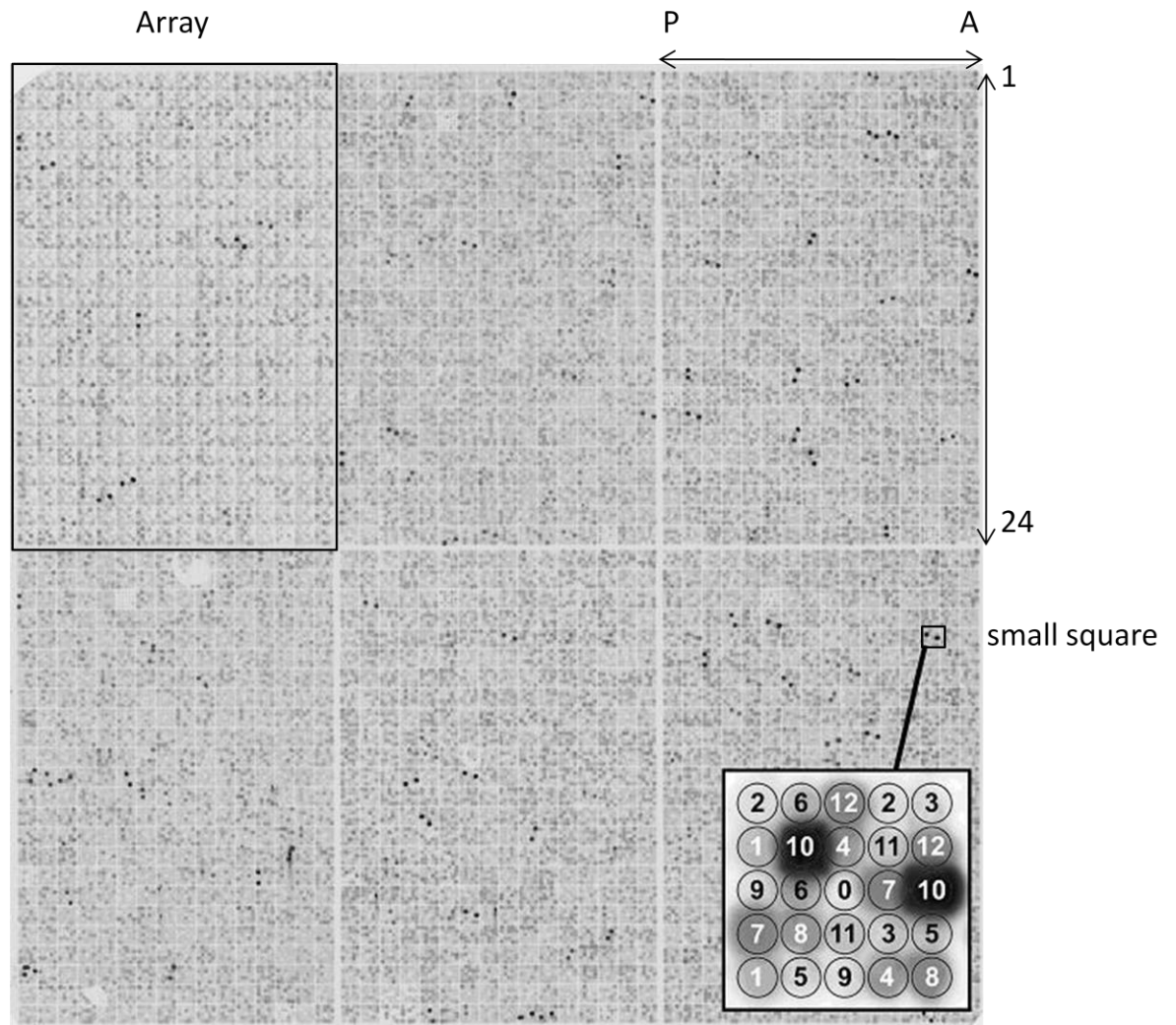


Figure 4.5. Identification of clones from gridded array.

Each of the 6 arrays represents 12 out of the 72 (384-well) plates. In each array, each small square represents gridding from the same well of 12 different 384-well plates. The plate pattern in the small squares is different for each of the 6 arrays. A close-up of one of the patterns in the small squares (plates 1-12) of the first array is shown. Each clone has been gridded in duplicate to avoid false positives during fluorescence analysis. In the close-up, a clone from plate 10 has a strong positive signal. To determine which well of the 384-well plate gave rise to this clone, each array has been gridded so that it has the same number-letter system as the 384-well plate itself (A-P on the horizontal axis, 1-24 on the vertical axis) – each square represents 1 of the 384 wells. In this case, plate 10 well C5 gave rise to the fluorescent clone.

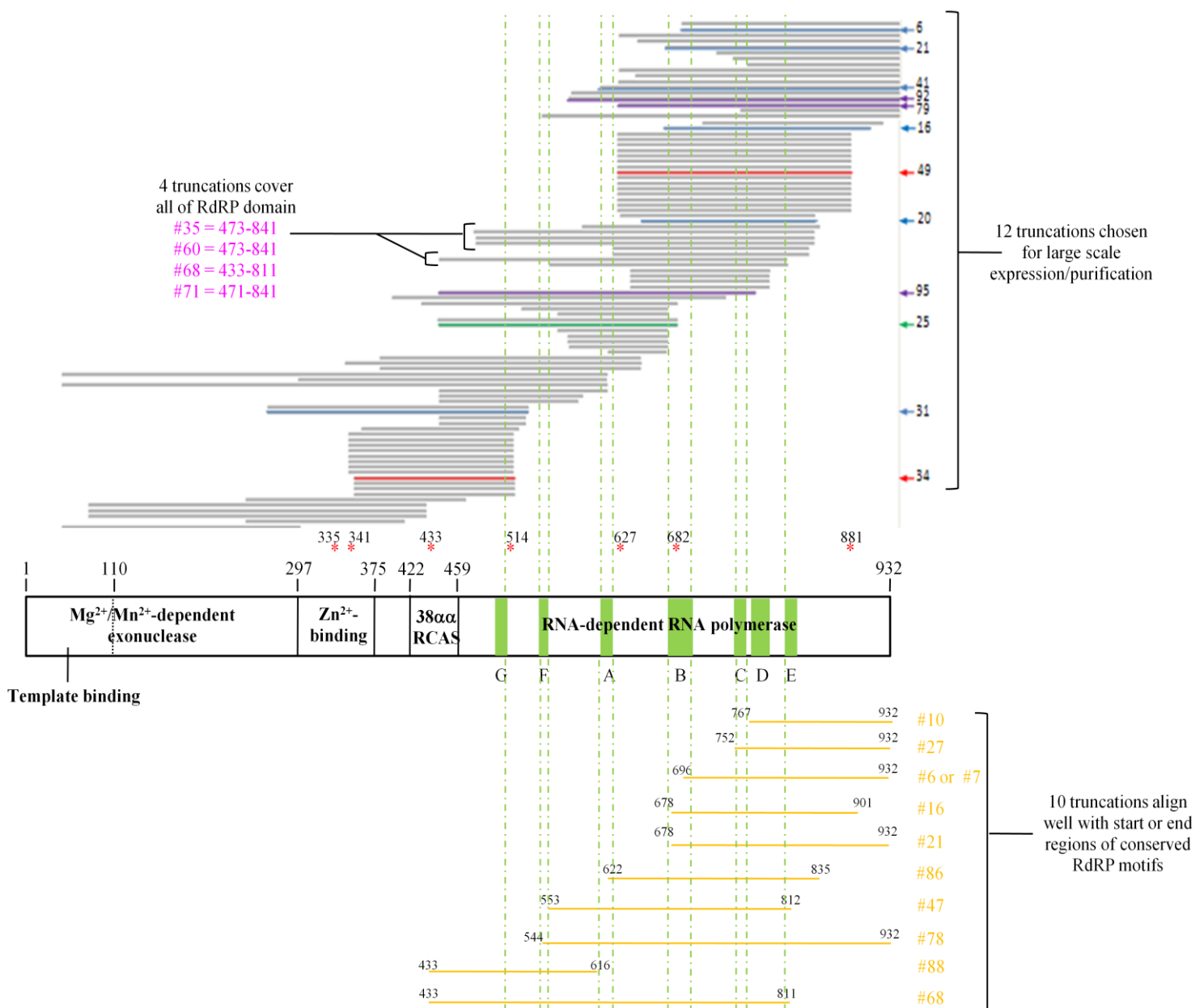
4.3. Analysis of the colony array

Signal quantification and image analysis were carried out using standard microarray software (Visual Grid, GPC-Biotech). The 27,648 clones were first ranked according to the fluorescence intensity of the His₆-tag signal and the numerical data exported to Excel (Microsoft). A threshold intensity was determined and used to identify positive hits over background signal. Clones below threshold intensity were considered not to have a His₆-tag and were thus removed from the dataset. The remaining clones were then ranked according to the fluorescence intensity of the streptavidin signal. A background level of biotinylation was seen in all colonies, so only colonies with a signal above background intensity were considered significant. The top (most intensely fluorescent) 95 clones were identified as described in **Fig. 4.5** and were recovered from glycerol stock plates and analysed by DNA sequencing (using T7 forward and T7 reverse primers), see **Fig. 4.6**.

The 95 hits ranged from 192 to 1788 bases, with the average being 731 bases (significantly smaller than the 1 – 2.796 kb expected). Many of these gene truncations cluster in certain regions along the length of NSP-12 varying by only a few amino acids. This clustering may indicate putative domain boundaries in NSP-12. Unfortunately, the clustering does not seem to support any of the models for NSP-12 proposed thus far (see Chapter 1, section 1.3.2 and **Fig. 1.17**). Hot-spots where many truncations start or end are highlighted in **Fig. 4.6**. Given the significant clustering around residues 433 and 841 spanning the putative RdRp domain, a synthetic construct (Syn) was designed to span this region and was scaled-up in parallel with other chosen gene truncations (see **Table 4.2**). Interestingly, none of the top 95 hits contained the first 20 residues at the N-terminus suggesting that truncations containing this region do not give rise to well-expressed correctly folded soluble protein in *E. coli*.

Figure 4.6. Sequencing and location of the top 95 hits along the length NSP-12 (932 $\alpha\alpha$).

Conserved motifs A-G (light green and extended with dotted lines) found in the RdRp sequence are described in (Bruenn, 2003) and their identification in the SARS RdRp sequence is shown in Xu *et al.*, (2003). Putative functional regions are indicated along the length of NSP-12 as described in Chapter 1, section 1.3.2 and **Fig. 1.17**. RCAS = replicative complex assembly sequence (Brockway *et al.*, 2003); * = hot spots, where many constructs start or end; 10 truncations align to start or end point of RdRp motifs (orange); 4 truncations span all RdRp motifs (magenta). 12 truncations were chosen for large scale expression/purification (see text) and are coloured as follows: most significant small scale expression results from Oxford and EMBL (blue); one of many constructs clustering in a particular region, but only see insoluble expression (red); covers part of putative RdRp domain, but only see insoluble expression (dark green); signs of solubility and covers most of putative RdRp domain (purple). Image generated using Vector NTI Invitrogen.



4.4. Small Scale Expression of ESPRIT constructs

The ESPRIT experiment indirectly identifies constructs that give rise to soluble protein expression using a biotin acceptor peptide (BAP tag) as a signal for solubility attached at the C-terminus (see Chapter 1, **Fig. 1.15**). To determine whether these constructs do express solubly in *E. coli* the following expression protocols were carried out. Note that protocols are different for work carried out at the EMBL in Grenoble (France) to those at the Oxford Protein Production Facility (OPPF) in Oxford (UK).

4.4.1. EMBL – Small Scale Expression

The top 95 gene truncations in *E. coli* BL21 RIL (DE3) cells were initially screened for expression at the EMBL. 40 μ L of starter culture (grown overnight in 1 mL LB) for each of the 95 truncations and a positive control (see section 4.4.2.), was transferred to a well of four 24-well blocks containing 4 mL of Terrific Broth (TB) supplemented with the 50 μ g/mL kanamycin. The 24-well blocks were placed in a HiGro Incubator (Gene Machines) at 37°C until an OD₆₀₀ of 0.6 was reached. Cells were then induced with 0.5 mM IPTG and grown for a further 16 hours at 25°C. Cell pellets were harvested by centrifugation at 6000 x *g* and resuspended in 4 mL of 10 mM Tris pH 8.0, 300 mM NaCl, 20% sucrose and 1 mg/mL lysozyme. After a 30 min incubation at 4°C, spheroblasts were isolated in the blocks by centrifugation at 6000 x *g* and were placed at -80°C for 20 min for one cycle of freeze-thaw to burst open the cells. Cells were then resuspended in 800 μ L lysis buffer [10 mM Tris pH 7.5, 0.5% Brij, 3 units/mL Benzonase (Merck; purity grade I, ≥ 25 U/ μ L)] supplemented with two EDTA-free protease inhibitor cocktail caplets (Roche). The lysate was transferred to a 96 well qiafilter plate (QIAGEN) containing 60 μ L of Ni-NTA resin (QIAGEN) per well and incubated at 4°C for 30 min. Purification was carried out on a robotic platform (Tecan),

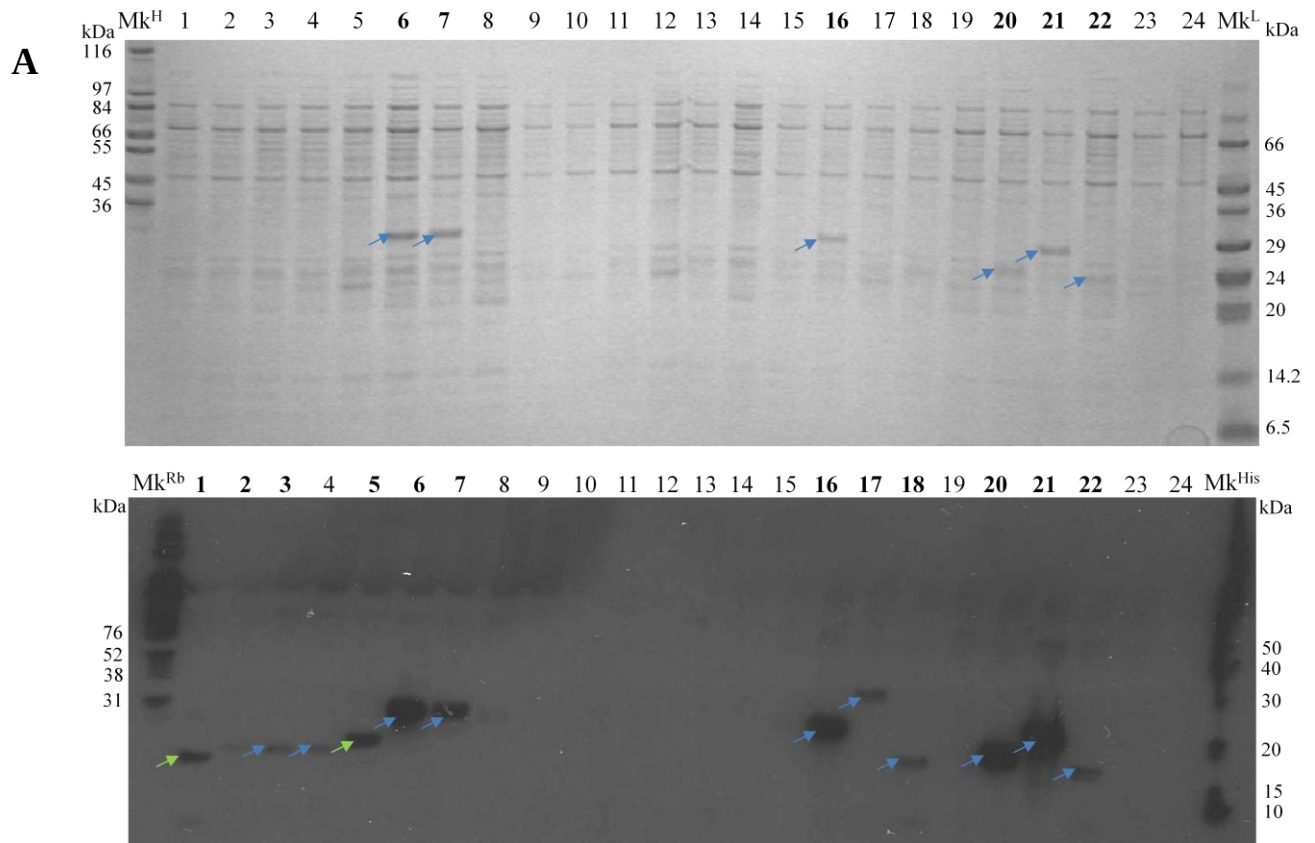
washing twice with wash buffer (10 mM Tris pH 7.5, 500 mM NaCl, 20 mM Imidazole) before eluting in 80 μ L of elution buffer (10 mM Tris pH 7.5, 500 mM NaCl, 500 mM Imidazole).

4.4.2. SDS-PAGE and Western Blot

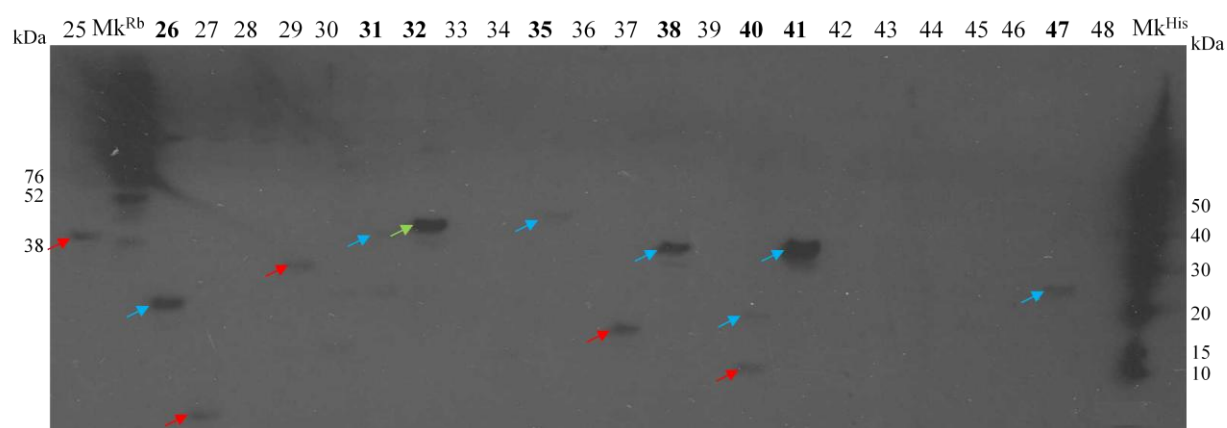
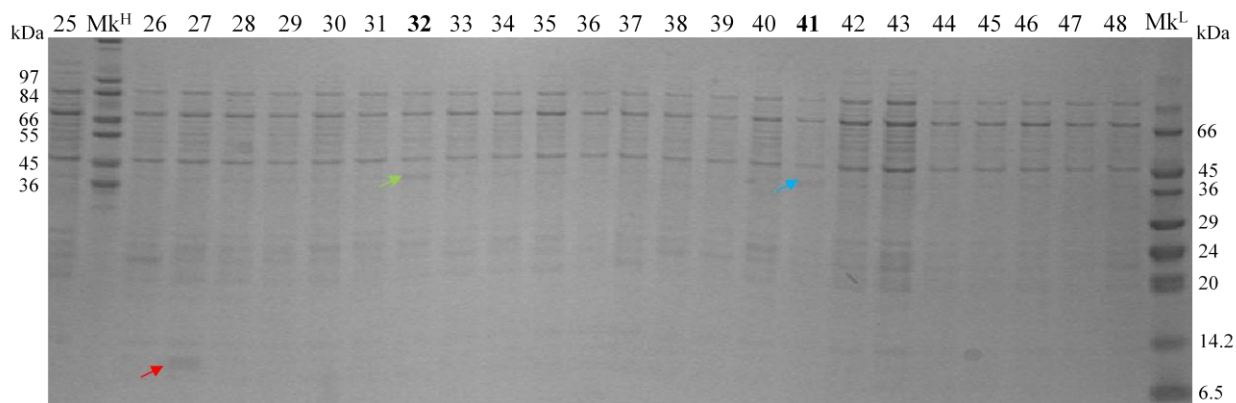
The SDS-PAGE gels and Western blots of the soluble fractions from the top 95 hits from the EMBL screen are shown in **Fig 4.7**. The positive control included at position 96 was mannose binding protein (MBP) with an N-terminal His₆ tag and C-terminal BAP tag provided by Darren Hart. SDS-PAGE was carried out as described in Chapter 2, section 2.1.4. Transfer of proteins from the SDS-PAGE gels onto nitrocellulose membranes for Western blotting experiments was carried out using the iBlot Dry Blotting System (Invitrogen) following the manufacturer's instructions. The nitrocellulose membrane was then blocked overnight in 30 mL blocking buffer (PBS-T with 5% casein milk powder) at 4°C, washed twice with 30 mL PBS-T for 5 min and finally washed again with 30 mL PBS without Tween-20 for a further 5 min. The membrane was then incubated for 1 h with the primary antibody (mouse anti-His₆ antibody, R&D systems) diluted 1:500 in 25 mL blocking buffer. The membrane was then washed with PBS-T and PBS as described previously before incubating with anti-mouse secondary antibody conjugated to horse radish peroxidase (HRP) enzyme (Sigma) diluted 1:5000 in 25 mL blocking buffer. The washing cycle was repeated again before visualising proteins by incubating with ECL-plus detection reagents (Amersham) following the manufacturer's instructions. Exposure was carried out in a dark room placing ECL hyperfilm (Amersham) over the membrane for approximately 15 s before developing.

Figure 4.7. EMBL – SDS-PAGE and Western blot of top 95 hits.

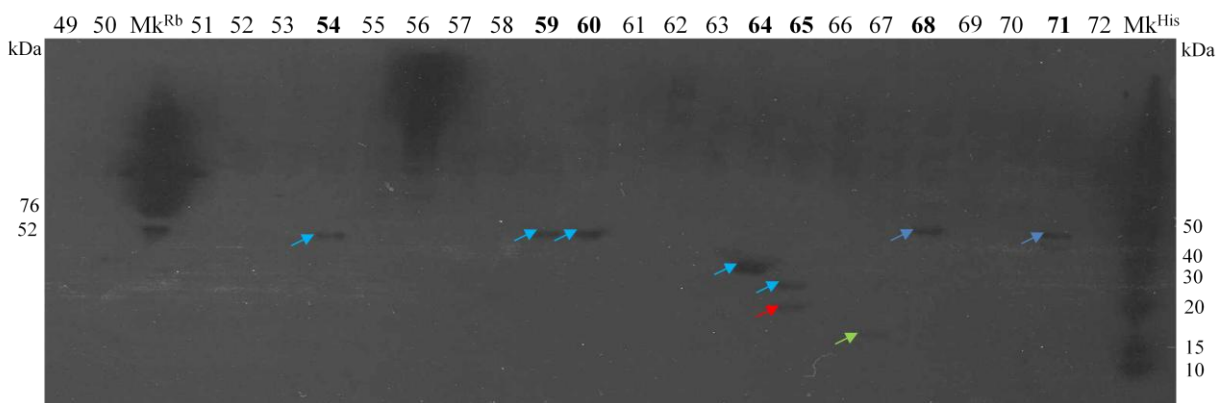
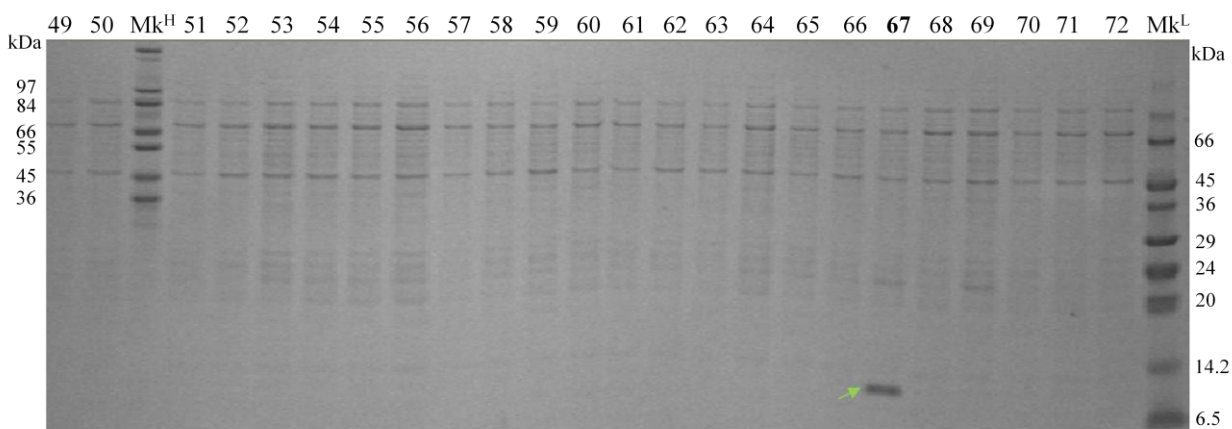
SDS-PAGE gels and Western blots for the EMBL small scale expression screen. Protein bands of interest are indicated by arrows coloured as follows: correct molecular mass from sequencing (blue); incorrect molecular mass from sequencing (red); similar molecular mass or insufficient sequencing data (light green); lane 96 positive control, His₆-MBP-BAP_{tag} (purple box; there is some overflow into lane 95). Protein ladders Mk^H and Mk^L are high and low range SigmaMarkers (Sigma) respectively; Mk^{Rbow} is ECL-Plex fluorescent rainbow marker (Amersham) used to visualise efficient transfer of proteins from gel to membrane; Mk^{His} is His-tagged BenchMark Protein Ladder (Invitrogen). The 96 samples were split into four sets: **(A)** Samples 1-24; **(B)** Samples 25-48; **(C)** Samples 49-72; **(D)** Samples 73-96. For **(A)-(D)** the SDS-PAGE gel of purified fractions is shown with the Western blot beneath it. Lanes are numbered and bold represents a result of interest.

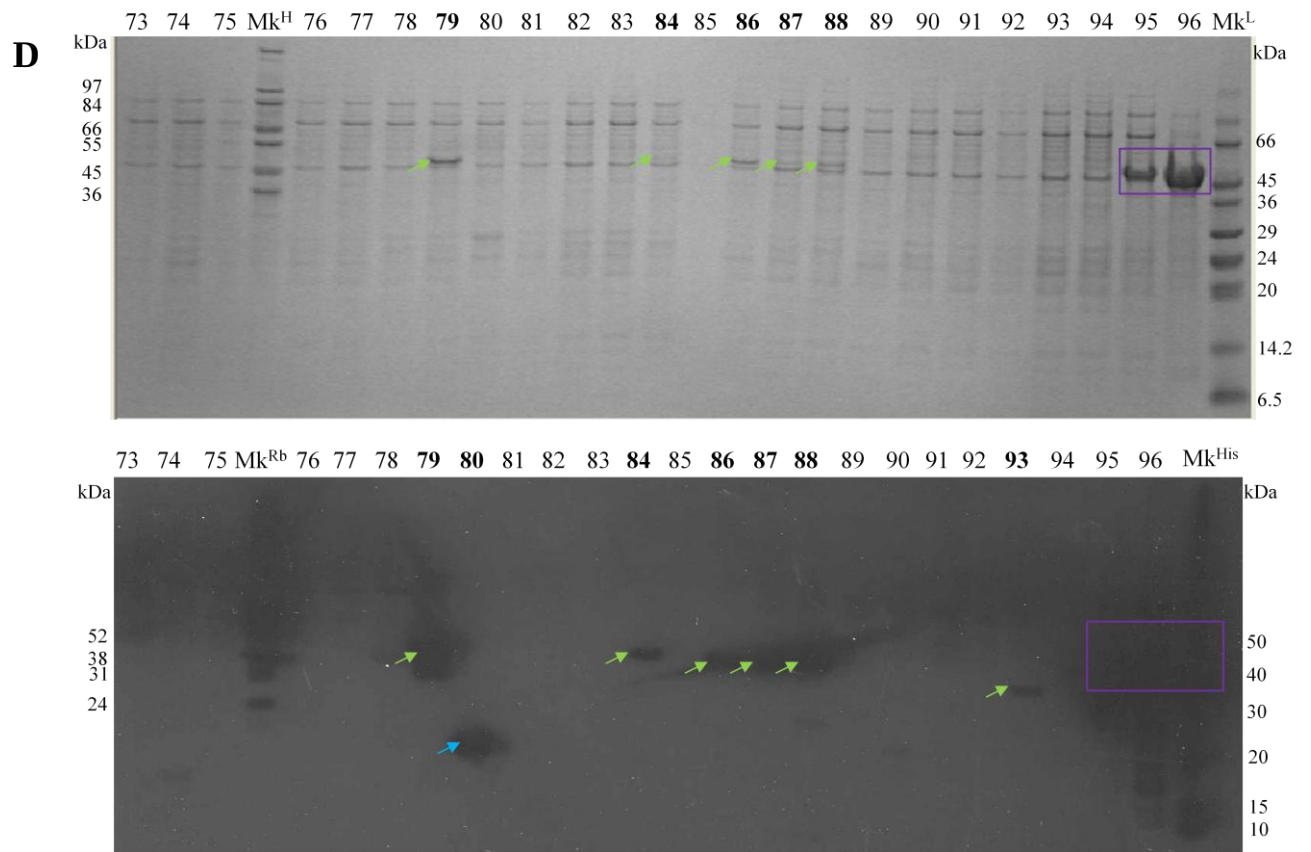


B



C



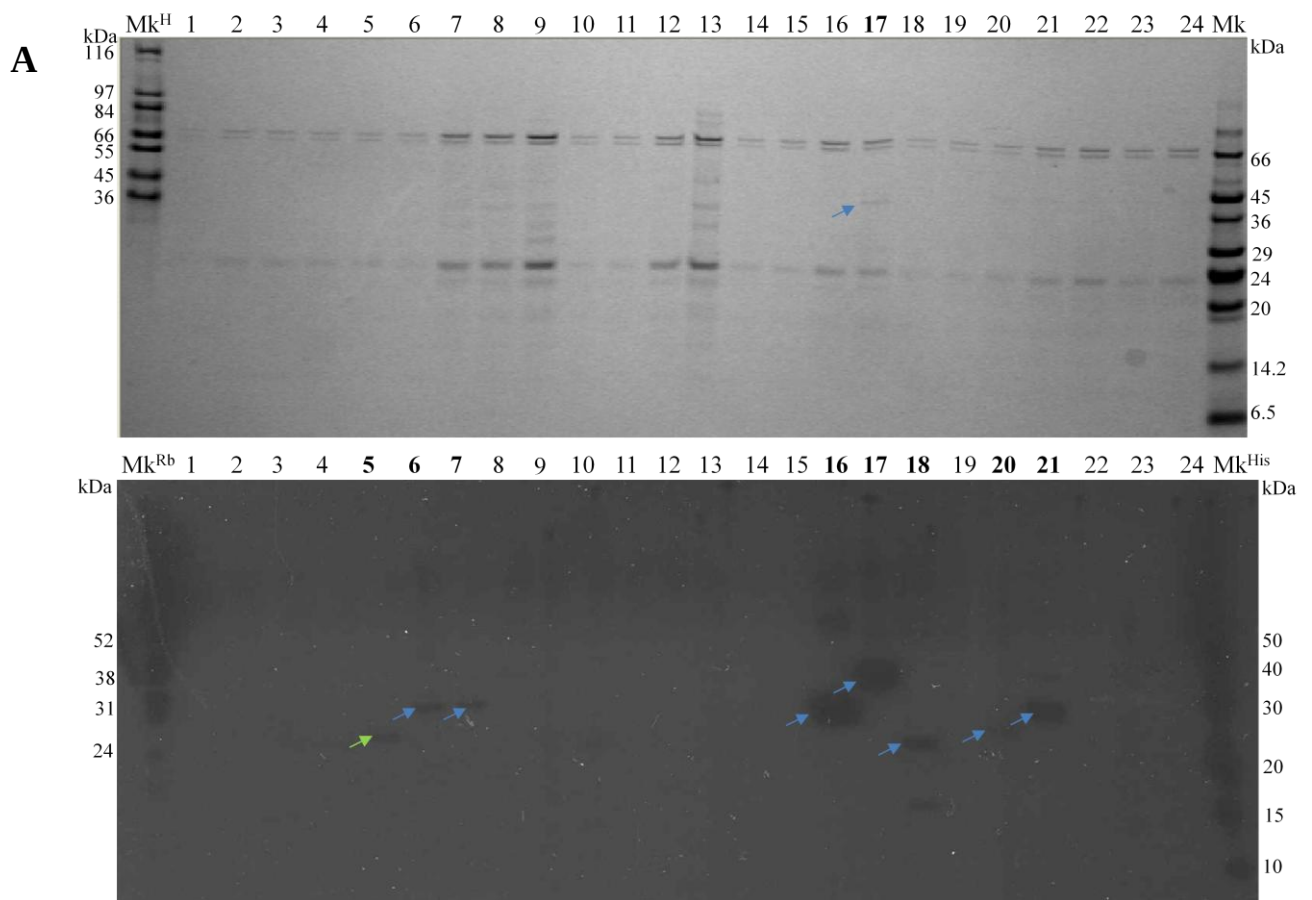


4.4.3. Oxford – Small Scale Expression

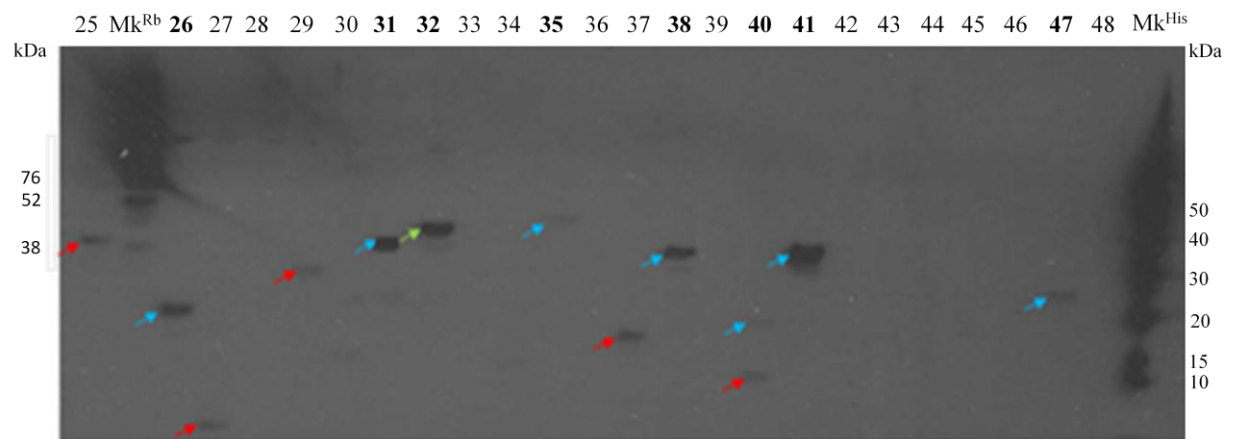
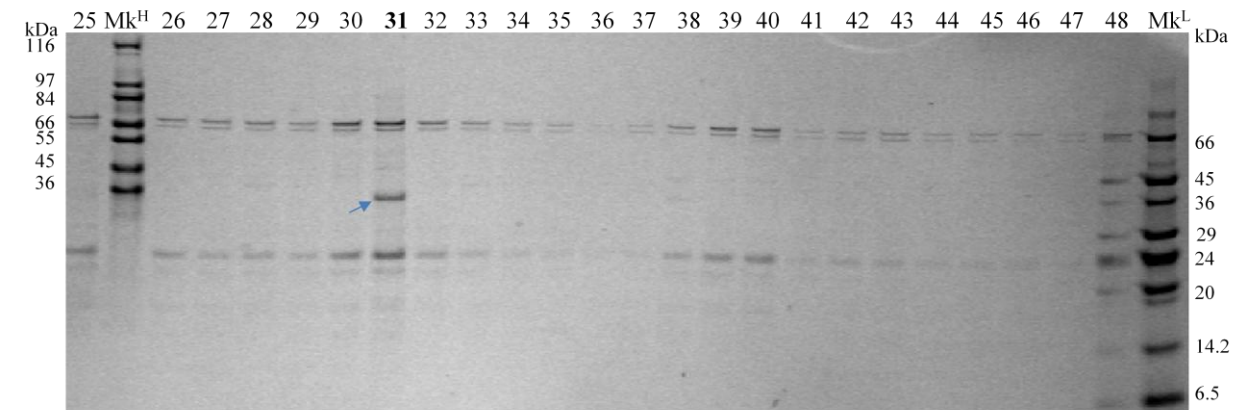
The top 95 constructs in *E. coli* BL21 RIL (DE3) cells were re-screened in Oxford following protocols reported in (Berrow *et al.*, 2007; Berrow *et al.*, 2006). In brief, 4 mL cultures in GS96 (QBioGene) or Power broth (Molecular dimensions) were grown at 37°C, until an OD₆₀₀ of 0.6 was reached. The temperature was then reduced to 20°C, IPTG added to a final concentration of 0.5 mM and the cultures grown for a further 18 h. A 1.5 mL aliquot of culture from each well was transferred into a 96-well deep-well block and harvested by centrifugation. Cell pellets were frozen at -80°C for a minimum of 1 h before screening for soluble expression on a robotic platform (BR8000) using standard QIAGEN Ni-NTA magnetic bead protocols. In this protocol, lysis is carried out by lysozyme and the soluble fraction separated from the insoluble fraction by centrifugation. Soluble fractions were analyzed by SDS-PAGE (described in Chapter 2, section 2.1.4.) and Western blot following the Invitrogen iBlot protocol described in section 4.4.2. (see **Fig. 4.8**; results are summarised in **Table 4.1**). Insoluble pellets were resuspended in solubilising solution (50 mM Tris pH 8.5, 300 mM NaCl, 8 M Urea, 30 mM Imidazole, 0.5% Triton-X and 1 mM TCEP). After centrifugation, solubilised protein containing a His₆-tag was isolated using Ni-Sepharose resin (GE Healthcare) and eluted with solubilising solution containing 300 mM Imidazole following the QIAGEN qiafilter plate protocol on a vacuum manifold. Insoluble (solubilised) fractions were also analyzed on SDS-PAGE gels (**Fig. 4.9** and summarised in **Table 4.1**).

Figure 4.8. Oxford – SDS-PAGE and Western blot of top 95 hits.

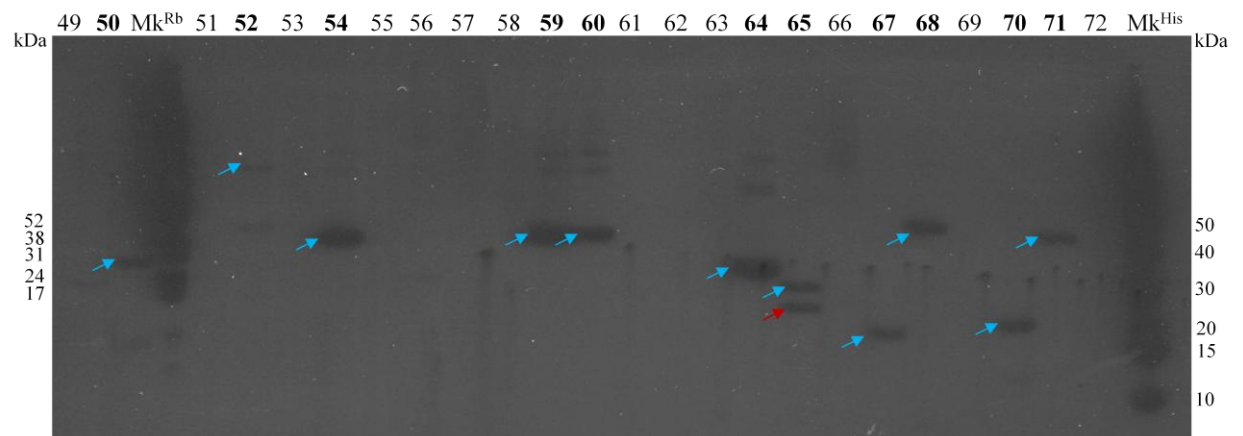
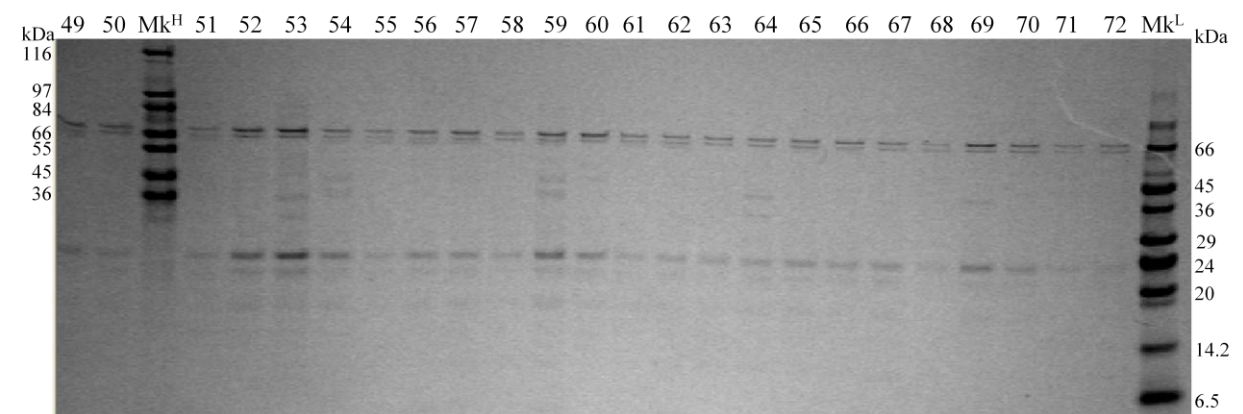
SDS-PAGE gels and Western blots for the Oxford small scale expression screen. Protein bands of interest are indicated by arrows coloured as follows: correct molecular mass from sequencing (blue); incorrect molecular mass from sequencing (red); similar molecular mass or insufficient sequencing data (light green). Protein ladders Mk^H and Mk^L are high and low range SigmaMarkers (Sigma) respectively; Mk^{Rbow} is ECL-Plex fluorescent rainbow marker (Amersham) used to visualise efficient protein transfer from gel to membrane; Mk^{His} is His-tagged BenchMark Protein Ladder (Invitrogen). The 96 samples were split into four sets: **(A)** Samples 1-24; **(B)** Samples 25-48; **(C)** Samples 49-72; **(D)** Samples 73-96. For **(A)-(D)** the resulting SDS-PAGE gel is at the top and Western blot result at the bottom. Lanes are numbered and bold represents a result of interest. No positive control was present in position 96 (empty lane).



B



C



D

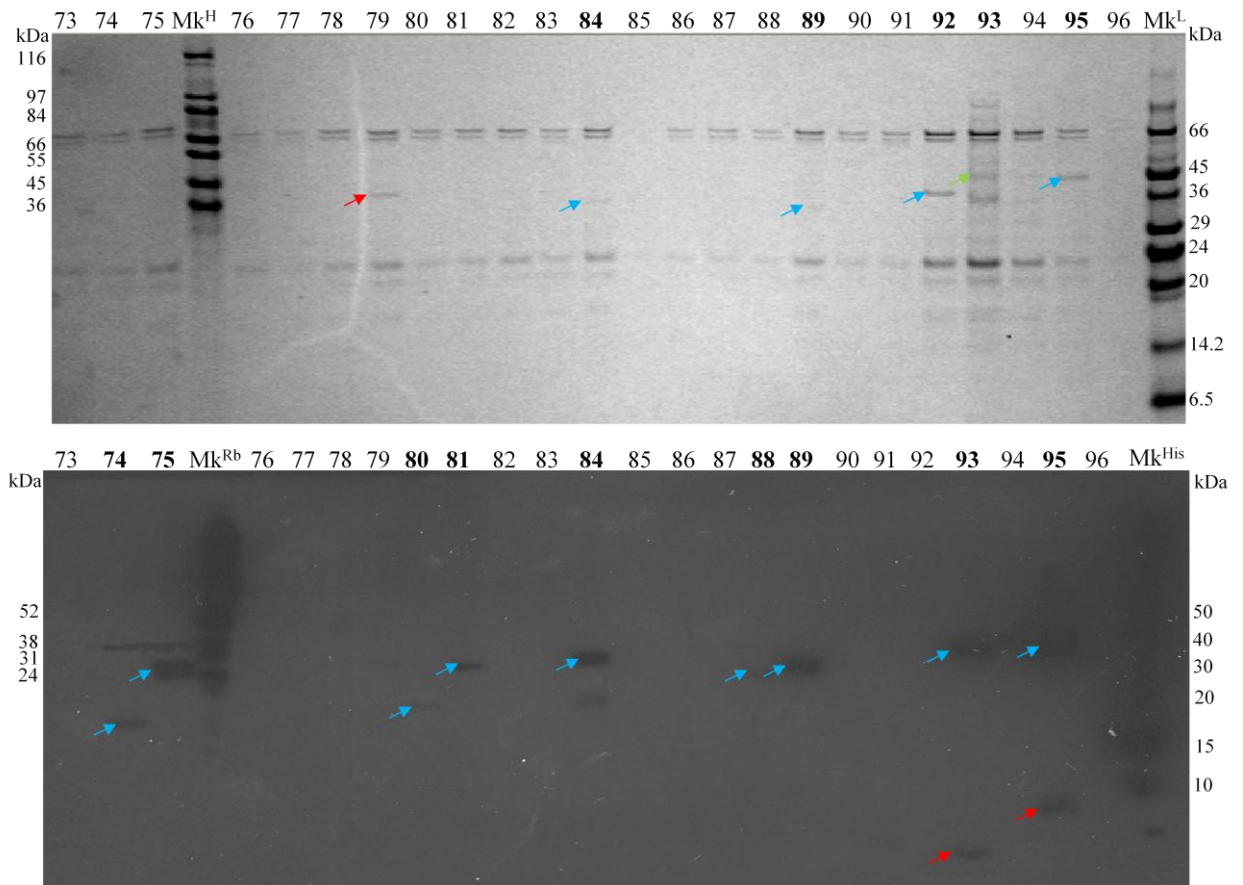
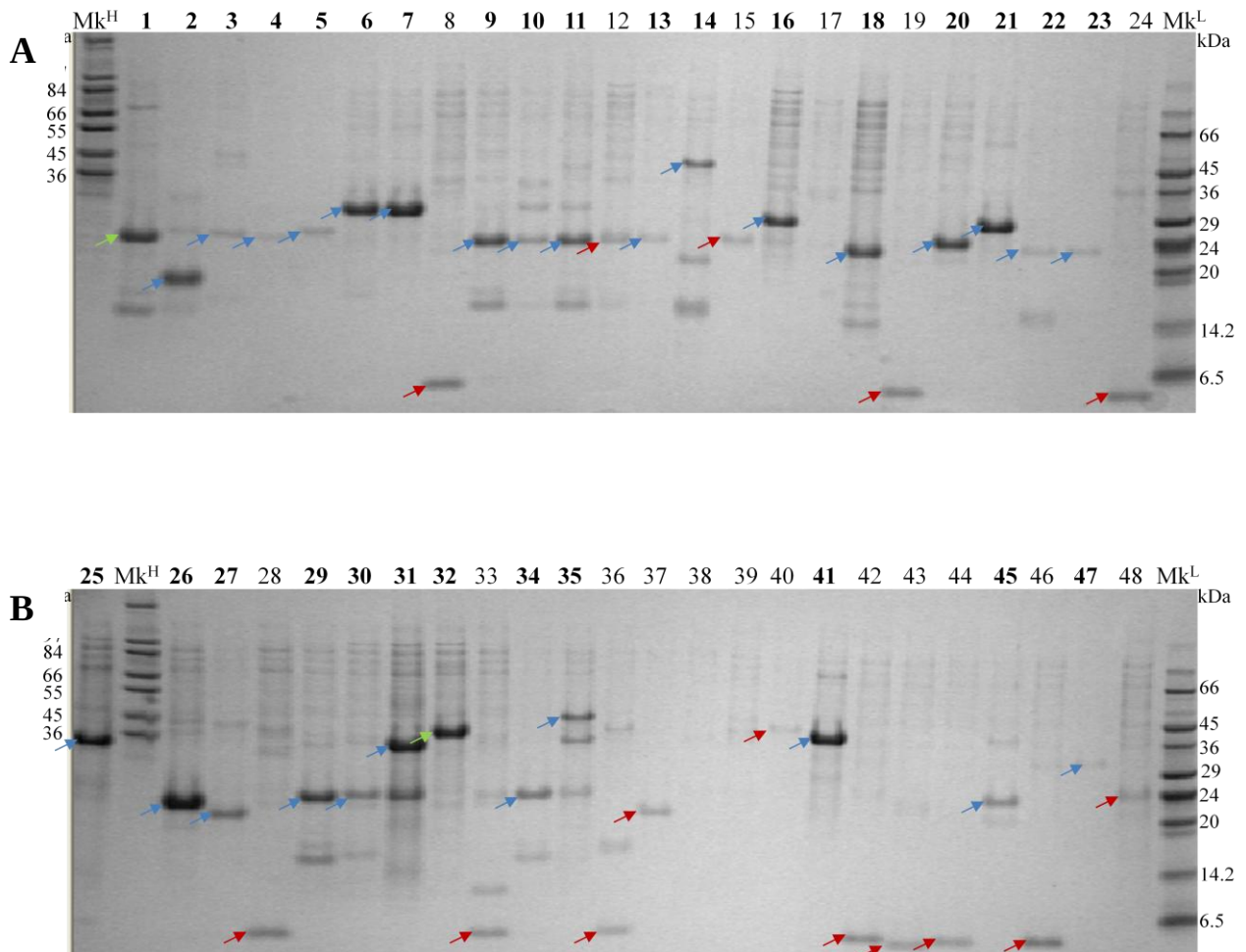
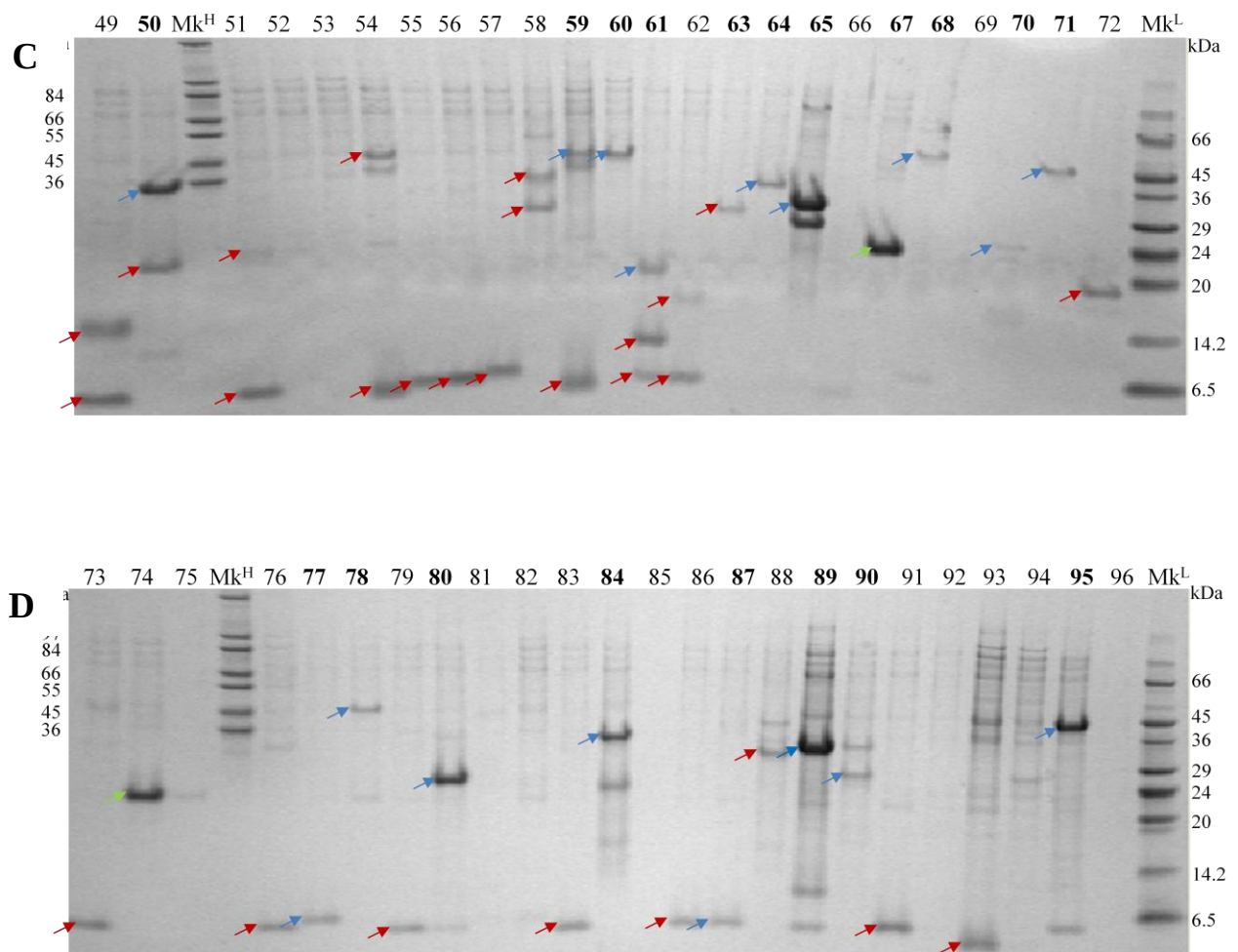


Figure 4.9. Oxford – Denaturing purification of top 95 constructs.

The 96 samples were split into four sets for SDS-PAGE analysis: **(A)** Samples 1-24; **(B)** Samples 25-48; **(C)** Samples 49-72; **(D)** Samples 73-96. Protein bands of interest are indicated by arrows coloured as follows: correct molecular mass from sequencing (blue); incorrect molecular mass from sequencing (red); similar molecular mass or insufficient sequencing data (light green). Protein markers were Mk^H and Mk^L for high and low range SigmaMarkers (Sigma) respectively. Lanes are numbered and bold represents a result of interest. A positive control was not included at position 96 (empty).





4.4.4. Small Scale Expression – Rescue Strategies

Unexpectedly, some of the results from the EMBL small scale purification do not agree with the screen repeated in Oxford (e.g. gene truncations #6 and #7 express solubly in the EMBL screen, but are only seen insolubly in the Oxford screen). Also, truncations that form clusters along the length of the NSP-12 from the sequencing data (e.g. #49, see **Fig. 4.6**), do not show any signs of soluble expression in either screen. From the denaturing purification many of these clustered gene truncations are represented by a low molecular mass band around 6.5 kDa on the SDS-PAGE gel (**Fig. 4.9**), suggesting that they are degraded upon expression.

In an attempt to identify whether changes in the small scale expression protocol would yield more soluble hits from the best 95, the following expression conditions were varied (data not shown):

- Media
 - Auto-induction with TB Overnight Express (Novagen)
 - IPTG-induction with GS96 (QBioGene) or with Power broth (Molecular dimension)
- Expression temperature: 15°C, 25°C and 37°C
- IPTG concentration: 0.5 mM, 1 mM or 2 mM
- Time given for expression before harvesting: 3 h, 6 h or 18 h post-induction

No new soluble hits were identified by these experiments. The most favoured condition was in fact the original Oxford condition (0.5 mM IPTG, Power broth and 18 h at 25°C). A summary table of the small scale expression results and details of all 95 clones is shown in **Table 4.1**.

Gene truncation	Start $\alpha\alpha$	End $\alpha\alpha$	Start sequence	End sequence	SDS-PAGE soluble	Western blot soluble	SDS-PAGE insoluble
1	382	745	AASGN	FVDEF	-	-	+++ Ox
2	573	682	QKLLK	GGTSS	-	-	+++ Ox
3	335	514	VDGVP	GKARL	-	+ EMBL	+ Ox
4	335	514	VDGVP	GKARL	-	+ EMBL	+ Ox
5	433	584	SSVEL	AATRG	-	++	+ Ox
6	696	932	ICQAV	HTVLQ	++ EMBL	+++	+++ Ox
7	696	932	ICQAV	HTVLQ	++ EMBL	++	+++ Ox
8	628	932	NMLRI	HTVLQ	-	-	-
9	341	515	VVSTG	KARLY	-	-	+++ Ox
10	767	932	NSNYA	HTVLQ	-	-	+ Ox
11	341	515	VVSTG	KARLY	-	-	+++ Ox
12	-	-	Vector	Only	-	-	-
13	335	514	VDGVP	GKARL	-	-	+ Ox
14	280	615	LCLFD	TPHLM	-	-	++ Ox
15	573	682	QKLLK	GGTSS	-	-	-
16	678	901	GGTSS	GHMLD	++ EMBL	+++	+++ Ox
17	648	932	LSHRR	HTVLQ	-	+	-
18	349	520	RELGV	DSSMS	-	+	+++ Ox
19	627	881	PNMLR	YADVF	-	-	-
20	653	842	YRLAN	LGAGC	++	+++	+++ Ox
21	678	932	GGTSS	HTLVQ	++ EMBL	+++	+++ Ox
22	335	514	VDGVP	GKARL	+ EMBL	+ EMBL	+ Ox
23	335	514	VDGVP	GKARL	-	-	+ Ox
24	627	881	PNMLR	YADVF	-	-	-
25	433	692	SSVEL	AYANS	-	-	+++ Ox
26	641	792	KHNTC	YQNNV	-	++	+++ Ox
27	752	932	HFSMM	HTVLQ	-	-	+ Ox
28	628	932	NMLRI	HTLVQ	-	-	-
29	341	515	VVSTG	KARLY	-	-	+++ Ox
30	335	514	VDGVP	GKARL	-	-	+ Ox
31	247	529	LTRAL	DALFA	++ Ox	+++	+++ Ox
32	734	932	NRDVD	HTVLQ	+ EMBL	++	+++ Ox
33	627	932	PNMLR	HTLVQ	-	-	-
34	341	515	VVSTG	KARLY	-	-	+ Ox
35	473	841	VEVVD	ILGAG	-	+	+ Ox
36	627	881	PNMLR	YADVF	-	-	-
37	-	-	Vector	Only	-	-	-
38	646	932	CNLSH	HTLVQ	-	+	-
39	617	680	WDYPK	KPGGT	-	-	-
40	335	514	VDGVP	GKARL	-	+	-
41	607	932	SDVET	HTVLQ	++	++	+++ Ox
42	627	881	PNMLR	YADVF	-	-	-

Gene truncation	Start $\alpha\alpha$	End $\alpha\alpha$	Start sequence	End sequence	SDS-PAGE soluble	Western blot soluble	SDS-PAGE insoluble
43	223	395	VQVAP	KRTTC	-	-	-
44	627	881	PNMLR	YADV F	-	-	-
45	523	682	DQDAL	GGTSS	-	-	+ Ox
46	627	881	PNMLR	YADV F	-	-	-
47	553	812	RARTV	GPHEF	-	+	+ Ox
48	331	653	RKIFV	SHRFY	-	-	-
49	627	881	PNMLR	YADV F	-	-	-
50	589	847	IGTSK	FVDDI	-	+ Ox	+++ Ox
51	627	881	PNMLR	YADV F	-	-	-
52	21	616	PCGTG	PHLMG	-	+ Ox	-
53	23	282	GTGTS	ERLCL	-	-	-
54	52	419	NCCRF	FNKDF	-	+	-
55	627	881	PNMLR	YADV F	-	-	-
56	627	881	PNMLR	YADV F	-	-	-
57	622	835	CDRAM	YPDPS	-	-	-
58	641	793	KHNTC	QNNVF	-	-	-
59	52	419	NCCRF	FNKDF	-	+	+ Ox
60	473	841	VEVVD	ILGAG	-	+	+ Ox
61	627	881	PNMLR	YADV F	-	-	+ Ox
62	627	881	PNMLR	YADV F	-	-	-
63	562	682	ICSTM	GGTSS	-	-	-
64	576	932	LKSIA	HTLVQ	-	+	+ Ox
65	223	462	VQVAP	YNLPT	-	+ Ox	+++ Ox
66	630	842	LRIMA	LGAGC	-	-	-
67	641	792	KHNTC	YQNNV	++ EMBL	+	+++ Ox
68	433	811	SSVEL	KGPHE	-	+	+ Ox
69	433	589	SSVEL	ATVVI	-	-	-
70	335	514	VDGVP	GKARL	-	+ Ox	+ Ox
71	471	841	FVVEV	ILGAG	-	+	+ Ox
72	575	682	LLKSI	GGTSS	-	-	-
73	627	881	PNMLR	YADV F	-	-	-
74	641	792	KHNTC	YQNNV	-	+ Ox	++ Ox
75	369	652	KELLV	LSHRF	-	++ Ox	-
76	627	881	PNMLR	YADV F	-	-	-
77	433	527	SSVEL	DQDAL	-	-	+ Ox
78	544	932	LKYAI	HTLVQ	-	-	+ Ox
79	627	932	PNMLR	HTLVQ	+	+ EMBL	-
80	720	916	VRNLQ	TSRYW	-	+++	+++ Ox
81	369	652	KELLV	LSHRF	-	+ Ox	-
82	21	616	PCGTG	PHLMG	-	-	-
83	627	881	PNMLR	YADV F	-	-	-
84	247	530	LTRAL	ALFAY	-	++	++ Ox
85	-	-	Vector	Only	-	-	-

Gene truncation	Start $\alpha\alpha$	End $\alpha\alpha$	Start sequence	End sequence	SDS-PAGE soluble	Western blot soluble	SDS-PAGE insoluble
86	622	835	CDRAM	YPDPS	++	++	-
87	433	527	SSVEL	DQDAL	-	++	-
88	433	616	SSVEL	PHLMG	++ ^{EMBL}	++	-
89	432	692	GSSVE	AYANS	+ ^{Ox}	+ ^{Ox}	+++ ^{Ox}
90	414	692	NFNKD	AYANS	-	-	+ ^{Ox}
91	760	932	DDAVV	HTLVQ	-	-	-
92	573	932	QKLLK	HTLVQ	+ ^{Ox}	-	-
93	52	419	NCCRF	FNKDF	-	+	-
94	562	682	ICSTM	GGTSS	-	-	-
95	433	777	SSVEL	QGLVA	+	+	+++ ^{Ox}
96	-	-	Positive	Control	+++	+++	n/a

Table 4.1. Summary Table of Small Scale Expression (EMBL and Oxford).

The abbreviations and symbols in the table refer to the following: ^{EMBL} = only EMBL screen; ^{Ox} = only Oxford screen; + = low protein expression of correct molecular mass; ++ = good protein expression of correct molecular mass; +++ = significant protein expression of correct molecular mass; - = no expression at correct molecular mass. The table is coloured in accordance with **Fig. 4.6**: most significant small scale expression results (blue); one of many constructs clustering in a particular region but only see insoluble expression (red); covers part of putative RdRp domain but only see insoluble expression (green); signs of solubility and covers most of putative RdRp domain (purple); span entire RdRp domain but only see insoluble expression (magenta); start or end at precise RdRp motif boundaries but only see insoluble expression (orange). The positive control (96) is His₆ tag-Maltose Binding Protein-BAP tag. Sequencing results for truncations 12, 37 and 85 identified only vector sequence and no sign of NSP-12 gene (these are indicated by *Vector only*, in the table). The start and end amino acid ($\alpha\alpha$) number is given for each gene truncation (full length NSP-12 RdRp is 932 $\alpha\alpha$) alongside the first and last five amino acids of the protein sequence.

4.5. Small Scale Expression – Analysis

4.5.1. Top soluble hits

Out of the 95 hits, 10 gene truncations show signs of soluble expression at small scale and are listed below (they are named by the first and last five amino acids and coloured in accordance with **Table 4.1** and **Fig. 4.6**):

- 1) ⁶⁹⁶ICQAV-HTVLQ⁹³², ⁶⁷⁸GGTSS-GHMLD⁹⁰¹ and ⁶⁷⁸GGTSS-HTVLQ⁹³²
 - Gene truncations: 6/7, 16 and 21
 - Strong soluble expression (EMBL screen) supported by strong bands in the Western blot
 - Weak solubility in Oxford screen (Western blot only), most in inclusion bodies

- 2) ⁶⁵³YRLAN-LGAGC⁸⁴², ⁶⁰⁷SDVET-HTVLQ⁹³² and ⁶²²CDRAM-YPDPS⁸³⁵
 - Gene truncations: 20, 41, and 86
 - Soluble expression in both EMBL and Oxford screens

- 3) ⁶⁴¹KHNTC-YQNNV⁷⁹² and ⁴³³SSVEL-PHLMG⁶¹⁶
 - Gene truncations: 67 and 88
 - Soluble expression (EMBL screen) supported by bands in the Western blot

- 4) ³¹LTRAL-DALFA²⁴⁷
 - Gene truncation: 31
 - Soluble expression (Oxford screen) supported by strong band in the Western blot

4.5.2. Clusters

Aligning the top 95 hits from the ESPRIT experiment along the length of NSP-12 revealed clusters of gene truncations in particular ‘hot spot’ regions (**Fig. 4.6**). In fact many of the clustered truncations are identical (same start and end point) and have been ‘selected for’ during the ESPRIT experiment as isolated domains. The six main clusters are listed below (they are named by the first and last five amino acids of the cluster and coloured where appropriate in accordance with **Table 4.1** and **Fig. 4.6**):

Cluster 1: ⁶²⁷PNMLR-YADV^{F881} & ⁶²⁷PNMLR-HTLVQ⁹³² (~18% of constructs)

- Gene truncations: 19, 24, 36, 42, 44, 46, 49, 51, 55, 56, 61, 62, 73, 76, 83 & 33/79
- All ⁶²⁷PNMLR constructs have a strong insoluble band at an incorrect molecular mass (consistently 6.5 kDa) likely due to degradation by proteases

Cluster 2: ³³⁵VDGVP-GKARL⁵¹⁴ (~8% of constructs)

- Gene truncations: 3, 4, 13, 22, 23, 30, 40, 70
- Weak insoluble band at the correct molecular mass; constructs have shown some signs of soluble expression in Western blots

Cluster 3: ³⁴¹VVSTG-KARLY⁵¹⁵ (~4% of constructs)

- Gene truncations: 9, 11, 29, 34
- Strong insoluble band at correct molecular mass

Cluster 4: ⁴³³SSVEL-..... (~8% of constructs)

- Gene truncations: 5, 25, 68, 69, 77, 87, 88, 95
- Most ⁴³³SSVEL constructs have a strong insoluble band at the correct molecular mass, although 88 does show signs of soluble expression in the EMBL screen that was not reproducible in Oxford

Cluster 5:- GGTSS⁶⁸² (~6% of constructs)

- Gene truncations: 2, 15, 45, 63, 72, 74 (also constructs 16 and 21 start at ⁶⁷⁸GGTSS⁶⁸²)
- GGTSS⁶⁸² constructs have a weak insoluble band at the correct molecular mass, but some show no expression at all

Cluster 6:- ILGAG⁸⁴¹ (~3% of constructs)

- Gene truncations: 35, 60 and 71
- ILGAG⁸⁴¹ constructs cover the entire putative RdRp domain, but have very low expression levels (weak bands in Western blot and insoluble protein screen)

Disappointingly none of the gene truncations starting and ending at the same point (forming clusters along the length of NSP-12), give rise to significant soluble protein expression. However, there is some evidence for soluble expression at the ⁴³³SSVEL and ⁶⁷⁸GGTSS boundaries that are also pulled out in the top soluble hits identified above (e.g. truncations #88, #16 and #21). Interestingly, ⁶⁷⁸GGTSS is the precise starting sequence of RdRp motif B for NSP-12 (see **Fig. 4.10**). A closer analysis of the truncations also identified #10, #27, #47, #68, #78 and #86 that align precisely with RdRp motifs (**Fig. 4.10**). This suggests that the ESPRIT experiment has successfully ‘trapped’ conserved regions that are essential to establish the scaffold for RdRp activity.

Figure 4.10. The conserved RdRp motifs A-G of NSP-12

The positions of conserved RdRp motifs A-G in SARS-CoV NSP-12 have been reported by Bruenn, (2003) and Xu *et al.*, (2003). Amino acids underlined and labelled represent gene truncations that start or end at or near putative RdRp motifs. Blue indicates a top soluble hit from small scale expression and orange identifies the other gene truncations covering part of a putative RdRp motif region (colouring is consistent with **Fig. 4.6**). Dashed lines indicate the position of gene truncation amino acids that are outside of the RdRp motifs.

Motif G: ⁴⁹⁹DKSAGFPFNKWGK⁵¹¹

Motif F: ⁵⁴⁴LKYAISAKNRARTVAGV⁵⁶⁰
 #78 #47

Motif A: ⁶¹²PHLMGWDYPKCDRAM⁶²⁶
 #88 #86

Motif B: ⁶⁷⁸GGTSSGDATTAYANSVFNICQAVTANVNALLST⁷¹⁰
 #16 or #21 #6 or #7

Motif C: ⁷⁵³FSMMILSDDAVVCYN⁷⁶⁷
 #27 #10

Motif D: ⁷⁷¹AAQGLVASIKNFKAVLYYQNNVFMSE⁷⁹⁶

Motif E: ⁸¹⁰HEFCSQHTMLV⁸²⁰
 #47 or #68

A summary of the expression results for the gene truncations chosen for large scale protein production is shown in **Table 4.2**. These constructs are also highlighted on the sequencing results (**Fig. 4.6**) using the same colour co-ordination and their positions along the length of NSP-12 are shown in detail in **Fig. 4.11**. The analysis of clustered gene truncations revealed that many truncations of different length start or end with residues ⁴³³SSVEL or ILGAG⁸⁴¹ immediately before and after the entire putative RdRp domain. A synthetic construct (Syn) was therefore designed to span this region (see re-cloning into pOPINF vector, section 4.6.2.1 and Appendix section A1.2.) and was scaled-up in parallel with the other chosen constructs.

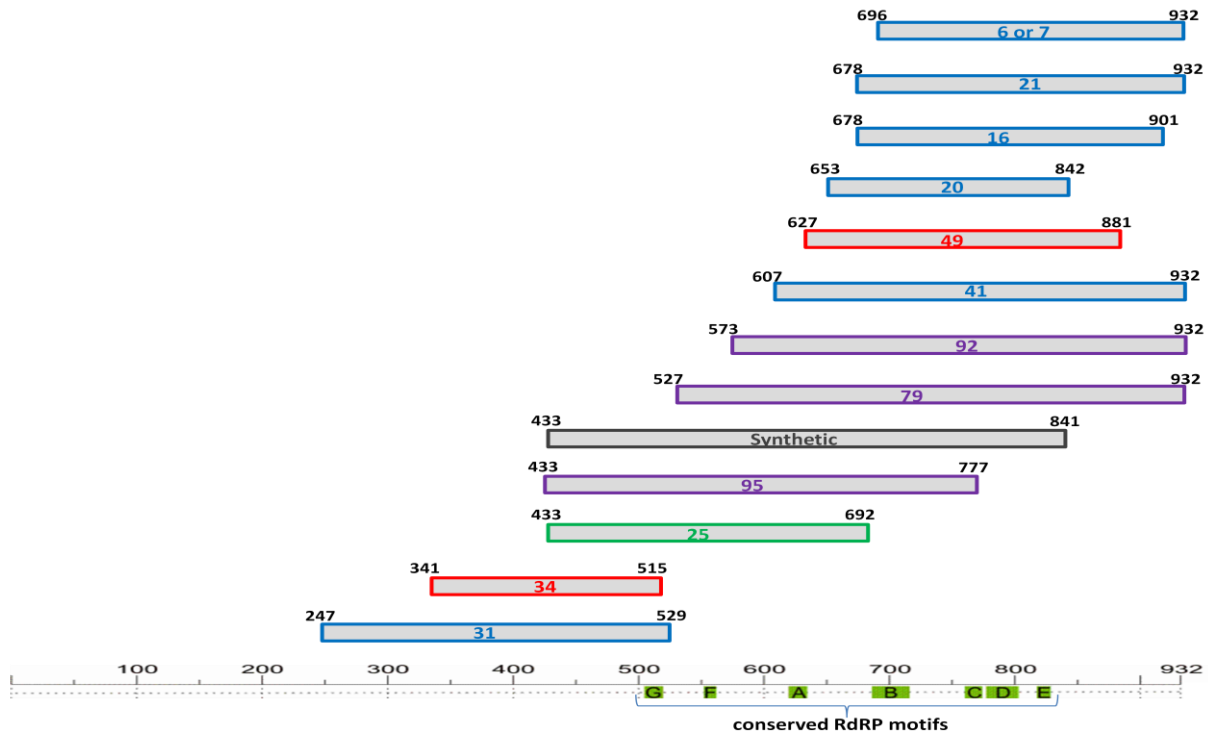
Table 4.2. Gene truncations chosen for large scale protein production.

The table summarises the small scale expression results for the 13 gene truncations: + = low protein expression; ++ = good protein expression; +++ = significant protein expression; — = no protein band at correct molecular mass; ^{EMBL} = only seen for EMBL small scale screen; ^{Ox} = only seen for Oxford small scale screen; Syn = synthetic construct (433-841) designed from sequencing data (see text and Appendix section A1.2.).

Gene trunc.	Start αα	End αα	SDS-PAGE soluble	Western blot soluble	SDS-PAGE insoluble	Reason why truncation was chosen for large scale screen
6	696	932	++ ^{EMBL}	+++	+++ ^{Ox}	Soluble hit supported by Western blot
16	678	901	++ ^{EMBL}	+++	+++ ^{Ox}	Soluble hit supported by Western blot
20	653	842	++	+++	+++ ^{Ox}	Soluble hit supported by Western blot
21	678	932	++ ^{EMBL}	+++	+++ ^{Ox}	Soluble hit supported by Western blot
25	433	692	—	—	+++ ^{Ox}	Insoluble, covers part of RdRp region
31	247	529	++ ^{Ox}	+++	+++ ^{Ox}	Soluble hit supported by Western blot
34	341	515	—	—	+ ^{Ox}	One of many hits clustered in this region
41	607	932	++	++	+++ ^{Ox}	Soluble hit supported by Western blot
49	627	881	—	—	—	One of many hits clustered in this region
79	627	932	+	++ ^{EMBL}	—	Signs of solubility, covers RdRp region
92	573	932	+ ^{Ox}	—	—	Signs of solubility, covers RdRp region
95	433	777	+	+	+++ ^{Ox}	Signs of solubility, covers RdRp region
Syn	433	841	n/a	n/a	n/a	Synthetic gene in pOPINF (section 4.6.2.1)

Figure 4.11. Positions of the gene truncations along the length of NSP-12 (932αα).

Conserved RdRp motifs are lettered and indicated in green. Gene truncations are coloured in accordance with **Table 4.2**.



4.6. Large Scale Protein Production

Protein was produced following published protocols (Berrow *et al.*, 2007; Berrow *et al.*, 2006). In brief, native proteins were produced in *E coli* BL21 RIL (DE3) cells in 500 mL cultures grown and induced as described in section 4.4.3. Cells were harvested by centrifugation and lysed using a Basic-Z cell disruptor (Constant systems) in 50 mM Tris pH 7.5, 500 mM NaCl, 20 mM Imidazole, 0.02% Tween-20. The His₆-tagged protein was purified by Nickel affinity (1 mL HisTrap) and size-exclusion chromatography (16/60 HiLoad Superdex 75 or 200) on an ÄKTExpress (GE healthcare). After loading the cleared lysate, the HisTrap column was washed to a UV baseline with 50 mM Tris pH 7.5, 500 mM NaCl, 20 mM Imidazole. The protein was eluted with 50 mM Tris pH 7.5, 500 mM NaCl, 500 mM Imidazole and loaded on to a pre-equilibrated (20 mM Tris pH 7.5, 200-500 mM NaCl) size exclusion column. In some cases, Nickel affinity and gel filtration were followed by a Q ion exchange step in an attempt to remove contaminants from the protein fragment of interest. Soluble fractions were analyzed by SDS-PAGE (described in Chapter 2, section 2.1.4.) and Western blot following the Invitrogen iBlot protocol described in section 4.4.2. Insoluble pellets were resuspended in solubilising solution (50 mM Tris pH 8.5, 300 mM NaCl, 8 M Urea, 30 mM Imidazole, 0.5% Triton-X and 1 mM TCEP) and then centrifuged for 30 min at 5000 x *g*. After centrifugation, solubilised protein containing a His₆-tag was isolated using Ni-Sepharose resin (GE Healthcare) and eluted with solubilising solution containing 300 mM Imidazole following the QIAGEN qiafilter plate protocol on a vacuum manifold. Insoluble (solubilised) fractions were also analyzed on SDS-PAGE gels. Protein concentration was measured using a NanoDrop ND-1000 Spectrophotometer (Thermo Scientific).

4.6.1. Large Scale Expression – Results

The results of the large scale expression screen are summarised in **Table 4.3**. The results were somewhat disappointing with only gene truncation #6 (⁶⁹⁶ICQAV-HTVLQ⁹³²), **Fig. 4.12. A and B**, showing very weak soluble expression and most ending up in inclusion bodies (see denaturing purification of the insoluble pellet, **Fig. 4.12 C**). All other truncations expressed from pHAR1119 are either not expressed or end up in inclusion bodies as insoluble aggregate (**Table 4.3**).

4.6.2. Large Scale – Rescue Strategies

In order to improve soluble expression of the 12 truncations from the small scale screen chosen for large scale protein production, a number of rescue strategies were attempted.

4.6.2.1. Re-cloning into pOPINF

The 12 truncations were re-cloned into pOPINF [a pTriEx-2 based vector containing a cleavable N-terminal His₆ tag with a human rhinovirus 3C protease site; see Berrow *et al.*, (2007) and Appendix section A1.1, **Table A1.1.** and **Fig. A1.1.** In addition, a synthetic construct (spanning residues 433-841), with start and end hot spot sequences (see section 4.3 and **Fig. 4.6.**), was PCR amplified from the codon-optimised SARS NSP-12 gene and also cloned into pOPINF (Appendix section A1.1, **Fig A1.1**). Importantly, the C-terminal BAP tag sequence provided by the pHAR1119 vector is lost during this re-cloning into pOPINF.

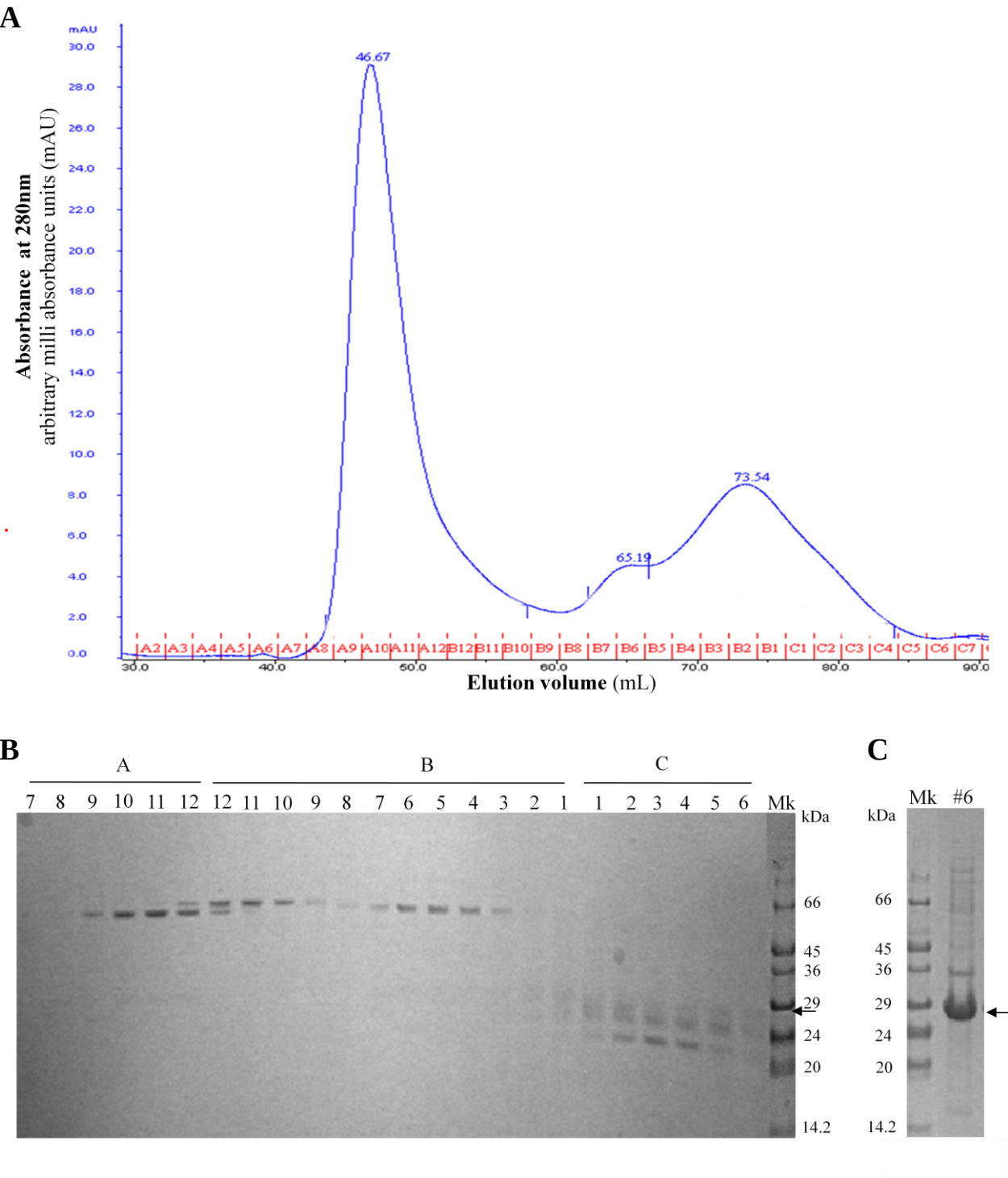
Table 4.3. Summary of large scale protein production results for truncations expressed from pHAR1119 and pOPINF.

Small scale expression results are provided in the table for comparison. Gene truncations are coloured as in **Table 4.2** and **Fig. 4.6**. The most significant result is indicated by a blue box. The expression of protein is scored as follows: + = faint protein band; ++ = significant protein band; +++ = strong protein band; * = confirmed by Western blot; – = no expression; n/a = experiment not carried out.

Gene trunc.	Vector	<i>E. coli</i> strain	Small scale expression (4 mL prep)		Large scale expression (0.5 L prep)	
			soluble	insoluble	soluble	insoluble
6	pHAR1119	BL21 (RIL)	+++*	+++	+*	+++
	pOPINF	Rosetta pLysS pRARE	n/a	n/a	–	++
16	pHAR1119	BL21 (RIL)	++	+++	–	++
	pOPINF	Rosetta pLysS pRARE	n/a	n/a	–	–
20	pHAR1119	BL21 (RIL)	+++*	+++	–	–
	pOPINF	Rosetta pLysS pRARE	n/a	n/a	–	–
21	pHAR1119	BL21 (RIL)	+++*	+++	–	–
	pOPINF	Rosetta pLysS pRARE	n/a	n/a	–	–
25	pHAR1119	BL21 (RIL)	–	+++	–	+++
	pOPINF	Rosetta pLysS pRARE	n/a	n/a	–	–
31	pHAR1119	BL21 (RIL)	+++*	+++	–	+++
	pOPINF	Rosetta pLysS pRARE	n/a	n/a	–	–
34	pHAR1119	BL21 (RIL)	–	++	–	++
	pOPINF	Rosetta pLysS pRARE	n/a	n/a	–	–
41	pHAR1119	BL21 (RIL)	+++*	+++	–	–
	pOPINF	Rosetta pLysS pRARE	n/a	n/a	–	–
49	pHAR1119	BL21 (RIL)	–	–	–	–
	pOPINF	Rosetta pLysS pRARE	n/a	n/a	–	–
79	pHAR1119	BL21 (RIL)	+++*	+++	–	–
	pOPINF	Rosetta pLysS pRARE	n/a	n/a	–	–
92	pHAR1119	BL21 (RIL)	+	–	–	–
	pOPINF	Rosetta pLysS pRARE	n/a	n/a	–	–
95	pHAR1119	BL21 (RIL)	+	+++	–	+++
	pOPINF	Rosetta pLysS pRARE	n/a	n/a	–	–
Syn 433-841	pOPINF	Rosetta pLysS pRARE	n/a	n/a	–	–

Figure 4.12. Large scale Nickel affinity-gel filtration of gene truncation #6 from pHAR1119 in BL21 RIL (DE3) cells.

(A) Elution profile (A_{280}) of proteins from gel filtration column. **(B)** SDS-PAGE gel of fractions eluted from gel filtration column, numbered in accordance with (A). **(C)** SDS-PAGE gel of denaturing purification of insoluble pellet. Mk = low range SigmaMarker (Sigma); arrows indicate the position of expressed truncation #6 (molecular mass: 32.5 kDa).



Cloning into pOPINF was carried out following published protocols (Berrow *et al.*, 2007) and used the In-Fusion enzyme ligation system (Clontech), see Appendix section A1.2. and **Fig. A1.2**. In brief, the PCR reaction was performed in 50 μ L reaction mixes using gene specific primers and KOD Hot Start DNA Polymerase according to the manufacturer's instructions (Novagen), with 30 pmol of each gene specific forward and reverse primers, appended with the In-Fusion extensions as appropriate (see Appendix section A1.1 and **Table A1.1**), and 1 μ L of pHAR1119 vector containing the gene truncation of interest (100 ng/ μ L) as template per reaction; for the synthetic construct the template was codon-optimised SARS NSP-12 gene. PCR products were purified using AMPure magnetic beads (Beckman-Coulter) according to the manufacturer's instructions and were eluted in 50 μ L of EB buffer (10 mM Tris-HCl, pH 8.0). 5 μ L of purified PCR product and 1 μ L of linearised pOPINF vector (100 ng/ μ L) were mixed in a well of an In-Fusion Dry-Down 96-well plate (Invitrogen) for each reaction and incubated at 42°C for 30 min. The In-Fusion reaction was quenched by diluting the mixture 5-fold with TE buffer (10 mM Tris-HCl, pH 8.0, 1 mM EDTA). 5 μ L of this quenched mixture was then used to transform OmniMaxII cells (Invitrogen) as described in section 4.1.2. The 13 new pOPINF constructs were then sequence-verified by GATC Biotech.

Large scale expression screens were carried out for the pOPINF constructs in the same way as detailed in section 4.6; see **Table 4.3** for a summary of the results. However, it is important to note that pOPINF constructs were transformed into *E. coli* Rosetta pLysS pRARE (DE3) cells (Novagen) containing additional plasmids encoding T7 lysozyme and rare tRNAs for arginine, leucine, and proline. Previously, *E. coli* BL21 RIL (DE3) cells were used when expressing from pHAR1119 vectors. *E. coli* BL21 RIL (DE3) cells do not

encode T7 lysozyme (pLysS) but have an additional plasmid encoding a different combination of rare tRNAs (arginine, isoleucine and leucine).

Unfortunately, no soluble or insoluble expression at large scale was seen for the truncations expressed from pOPINF in Rosetta pLysS pRARE (DE3) cells (**Table 4.3**). When expressed from the pHAR1119 vector at large scale 6 out of the 12 original truncations showed strong signs of insoluble expression. This difference suggests that expression of truncations is sensitive to the choice of *E. coli* cell line or equally that the C-terminally biotinylated BAP tag provided by the pHAR1119 vector has a positive effect on expressability of these proteins in *E. coli*.

4.6.2.2. Different *E. coli* strains

In an attempt to increase soluble protein expression, gene truncations #6 and #31 from pOPINF and pHAR1119 were transformed into *E. coli* B834 (DE3) methionine auxotroph cells (Novagen) and *E. coli* C41 (DE3) cells (Lucigen) following the procedure described in section 4.1.2. *E. coli* C41 (DE3) cells demonstrate a higher resistance to apoptosis when expressing toxic proteins (Dumon-Seignovert *et al.*, 2004), see **Fig. 4.13** for an example SDS-PAGE gel and Western blot for small scale expression in *E. coli* C41 (DE3) cells. Truncations #6 and #31 from pOPINF were also transformed into *E. coli* BL21 RIL (DE3) cells to determine whether loss of expression from re-cloning was due to the Rosetta pLys pRARE cell line. Soluble fractions were analyzed by SDS-PAGE (described in Chapter 2, section 2.1.4.) and Western blot following the Invitrogen iBlot protocol described in section 4.4.2. Insoluble (solubilised) fractions were also analyzed on SDS-PAGE as in section 4.4.3 and any hits were scaled-up as in section 4.6. A summary of the results is shown in **Table 4.4**.

Figure 4.13. Example of small scale expression for truncations #6 and #31 from pOPINF and pHAR1119 in *E. coli* C41 cells.

(A) SDS-PAGE gel of soluble fractions; Mk = low range SigmaMarker (Sigma).

(B) Western blot of soluble fractions; Mk^{His} = His-tagged BenchMark Protein Ladder (Invitrogen).

(C) SDS-PAGE gel of pHAR1119 fractions from denaturing purification of insoluble pellet.

Arrows indicate the positions of truncations #6 and #31 expressed at the correct molecular mass. For pOPINF, truncations #6 and #31 have a molecular mass of 29.7 kDa, and 34.5 kDa respectively; for pHAR1119, truncations #6 and #31 have a molecular mass of 32.5 kDa and 37.3 kDa respectively.

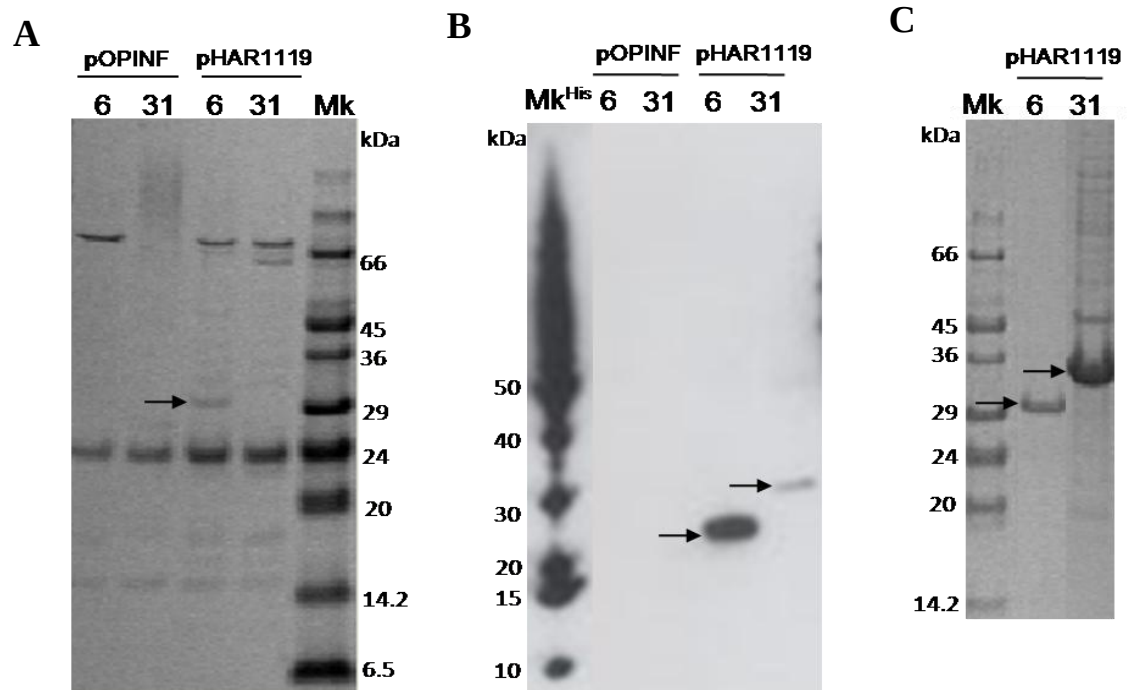


Table 4.4. A summary of small and large scale *E. coli* cell line expression results.

Blue box indicates the most significant results. Protein expression is scored as follows: + = weak protein expression; ++ = good protein expression; +++ = significant protein expression; * = confirmed by Western blot; – = no expression; n/a = experiment not carried out.

Gene trunc.	Vector	<i>E. coli</i> strain	Small scale expression (4 mL prep)		Large scale expression (0.5 L prep)	
			soluble	insoluble	soluble	insoluble
6	pHAR1119	B834	+	+	*	+
	pHAR1119	C41	+	++	+	+
	pOPINF	B834	–	–	n/a	n/a
	pOPINF	C41	–	–	n/a	n/a
	pOPINF	BL21 RIL	–	–	n/a	n/a
31	pHAR1119	B834	–	–	–	–
	pHAR1119	C41	*	+++	–	+
	pOPINF	B834	–	–	n/a	n/a
	pOPINF	C41	–	–	n/a	n/a
	pOPINF	BL21 RIL	–	–	n/a	n/a

No soluble or insoluble expression was seen for truncations #6 or #31 in *E. coli* BL21 RIL (DE3) cells suggesting that the cell line was not the reason for the change in expression levels and implies that expression from pOPINF is much less favourable than from pHAR1119. In fact, soluble expression was only seen for pHAR1119 constructs in all cell lines tested. These data suggest that biotinylation on the BAP tag (provided by the pHAR1119 vector) had a positive effect on the solubility and indeed expressability of constructs.

The best hit as large scale was #6 (pHAR1119) expressed in *E. coli* C41 (DE3) cells. Large scale purification of this condition is shown in **Fig. 4.14**. Four protein peaks are seen from the gel filtration elution profile, however evidence for protein #6 is present in all fractions of the resulting SDS-PAGE gel (**Fig. 4.14, B**). A Western blot of the eluted fractions (**Fig. 4.14, C**) revealed that His₆-tagged protein eluted in the first few fractions coming off the gel filtration column suggesting that the protein was in an aggregated or unfolded state.

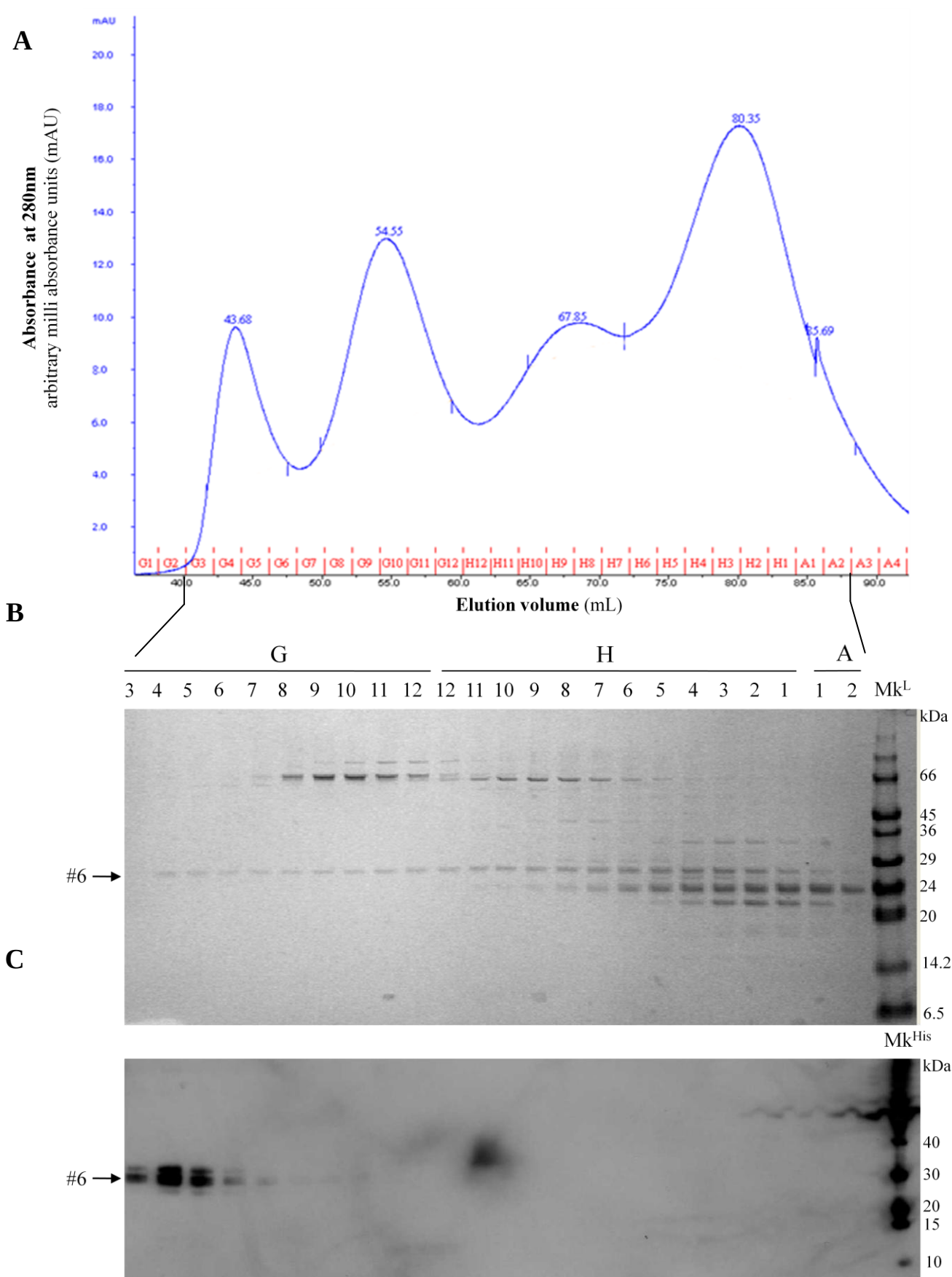
Figure 4.14 (overleaf). Example large scale Nickel affinity purification followed by gel filtration for truncation #6 (pHAR1119) expressed in C41 (DE3) cells.

(A) Elution profile (A_{280}) of proteins from gel filtration column.

(B) SDS-PAGE gel of eluted fractions (numbered) from gel filtration column; Mk^L = low range SigmaMarker (Sigma).

(C) Western blot of eluted fractions, Mk^{His} = His-tagged BenchMark Protein Ladder (Invitrogen).

An arrow indicates the position of truncation #6 (molecular mass is 32.5 kDa when expressed from pHAR1119), it runs slightly lower on the SDS-PAGE gel as seen in **Fig. 4.13**.



4.6.2.3. *Fusion protein rescue strategy*

The three best soluble hits from small scale expression (#6, #16 and #31) and the synthetic construct (residues 433-841) were re-cloned as protein fusions with SUMO (Small ubiquitin-like modifier) or with GST (Glutathione-S-transferase with a 3C protease cleavage site) in vectors pOPINS and pOPINJ respectively that provide the chosen fusion tag (see Appendix section A1.1 and **Table A1.1.**); the re-cloning was carried out following published protocols (Berrow *et al.*, 2007) and the In-Fusion ligation system described in section 4.6.2.1. Fusion proteins were expression-tested at large scale as in section 4.6. SUMO and GST tags could be cleaved using His₆-tagged SUMO protease or His₆-tagged 3C protease respectively. **Table 4.5** summarises the experimental results for the fusion protein constructs.

Unfortunately, soluble expression of fusion protein was not seen for any of the chosen 13 constructs re-cloned into pOPINS and pOPINJ. In fact only four fusion proteins showed any signs of expression, ending up in inclusion bodies as identified by denaturing purifications of the insoluble pellets following the protocol described in section 4.6 (see **Fig. 4.15**). Insoluble expression is seen for both SUMO- and GST-tagged synthetic construct containing ‘hot spot’ starting sequences (see **Fig. 4.6**) suggesting that these boundaries are more favourably expressed as expected.

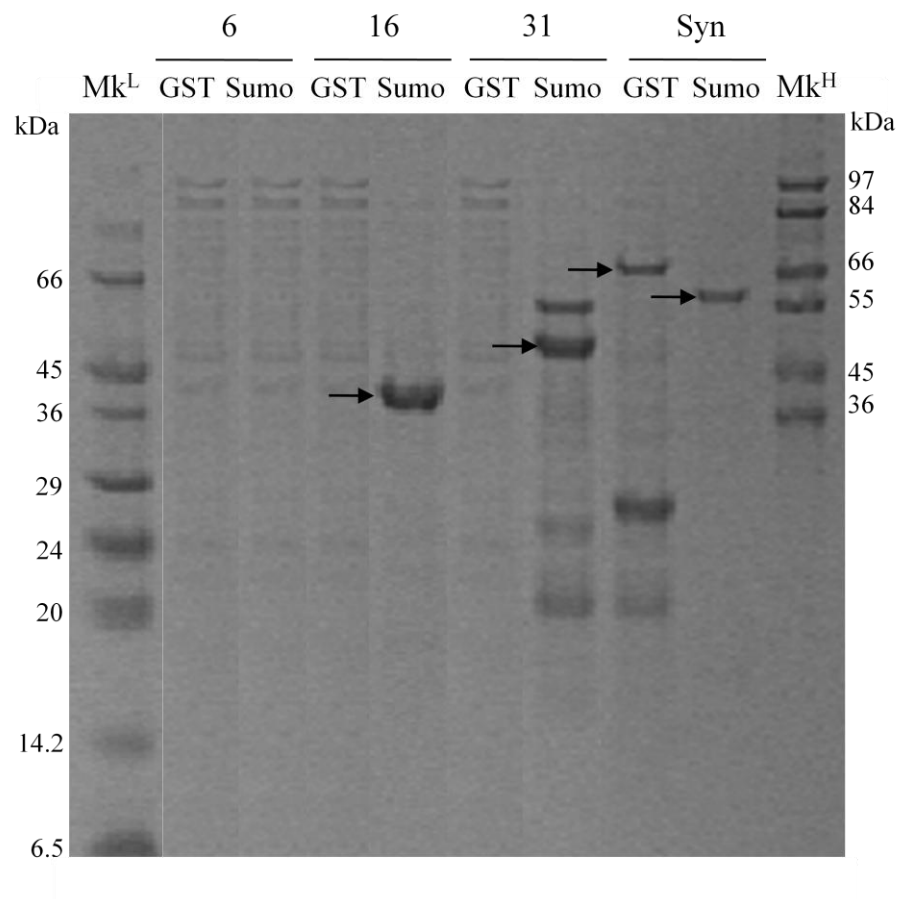
Table 4.5. Large scale fusion protein results.

Molecular masses (kDa) for the resulting fusion proteins (gene truncation plus fusion tag) are listed in the table alongside the large scale expression results for each fusion graded as follows: ++ = good protein expression; +++ = significant protein expression; – = no expression. N-His-GST indicates gene fusions containing an N-terminal His₆-tag followed by glutathione-S-transferase protein. N-His-SUMO indicates fusions containing an N-terminal His₆-tag followed by small ubiquitin-like modifier protein. Molecular masses in bold reflect fusions that could be expressed (insolubly) in *E. coli* Rosetta pLysS pRARE cells, see Fig. 4.15.

Gene trunc.	Vector	Tag	Molecular mass (kDa)	<i>E. coli</i> strain	Large scale expression (0.5 L prep)	
					soluble	insoluble
6	pOPINJ	N-His-GST	58.7	Rosetta pLysS pRARE	–	–
	pOPINS	N-His-SUMO	42.7	Rosetta pLysS pRARE	–	–
16	pOPINJ	N-His-GST	56.6	Rosetta pLysS pRARE	–	–
	pOPINS	N-His-SUMO	40.6	Rosetta pLysS pRARE	–	+++
31	pOPINJ	N-His-GST	63.5	Rosetta pLysS pRARE	–	–
	pOPINS	N-His-SUMO	47.5	Rosetta pLysS pRARE	–	+++
Syn	pOPINJ	N-His-GST	75.4	Rosetta pLysS pRARE	–	++
	pOPINS	N-His-SUMO	59.4	Rosetta pLysS pRARE	–	++

Figure 4.15. SDS-PAGE gel from denaturing purification of insoluble pellets.

Insoluble (solubilised) fusions were analysed by SDS-PAGE to assess whether lack of fusion proteins in soluble fractions was a result of protein aggregation in inclusion bodies upon expression in *E. coli*. Arrows indicate the positions of fusion proteins expressed with the correct molecular mass (see **Table 4.5**). Mk^L and Mk^H are high and low range SigmaMarkers (Sigma) respectively (masses are listed in kDa).



4.6.2.4. Refolding strategy

Many of the gene truncations could be expressed in *E. coli* but ended up in inclusion bodies rather than as correctly folded soluble protein domains. As a final attempt to obtain soluble protein, refolding from inclusion bodies was carried out for truncations #6, #25, #31 and #95 expressed from pHAR1119 in *E. coli* BL21 (DE3) RIL cells that yielded the most significant insoluble expression at large scale (see **Table 4.3.**).

Firstly, gene truncations #6, #25, #31 and #95 were expressed at large scale as described in section 4.6 but using 2 L of culture (rather than 0.5 L) in order to obtain a larger cell pellet. The pellet was resuspended in 1x PBS and lysed using a Basic-Z cell disruptor (Constant systems) as described in section 4.6. Crude inclusion bodies were then pelleted at 12,000 x g for 30 min and the supernatant discarded. The pellet was resuspended in 30 mL Triton wash buffer (0.5% Triton X100, 50 mM Tris-HCl, pH 8.0, 100 mM NaCl, 0.1% sodium azide) using a homogeniser to fully grind the pellet, and centrifuged again (25,000 x g for 10 min) to pellet the inclusion bodies. The Triton wash/centrifugation step was then repeated four times. After this the pellet was resuspended for a final time but in 50 mM Tris-HCl, pH 8.0, 100 mM NaCl to remove Triton and then pelleted at 25,000 x g for 10 min. The purified inclusion bodies were dissolved in guanidium hydrochloride (GnHCl) buffer (6 M GnHCl, 50 mM Tris-HCl, pH 8.5, 100 mM NaCl, 10 mM EDTA and 10 mM DTT) for 1 h at 4°C before centrifuging at 25,000 x g for 20 min. The protein concentration of the supernatant (denatured proteins) was then assessed using a NanoDrop ND-1000 Spectrophotometer (Thermo Scientific) and a sample was analysed by SDS-PAGE by diluting 25-fold in SDS loading buffer (2X stock, 100 mM Tris pH 6.4, 4% w/v SDS, 0.2 % w/v Bromophenol blue, 20 % v/v glycerol) and following the procedure outlined in Chapter 2, section 2.1.4. Supernatant was then stored in ~100 mg aliquots at -20°C.

Refolding was carried out by rapid dilution. A single 100 mg aliquot of denatured protein was defrosted and added drop-wise into 1 L of refold buffer (200 mM Tris-HCl, pH 8.0, 200 mM NaCl, 10 mM EDTA, 1 M L-arginine, 0.1 mM PMSF, 2 mM reduced glutathione, 0.2 mM oxidised glutathione) to a final concentration of 0.1 mg/mL in a 5 L flask at 4°C using a magnetic stirrer to fully mix the solution. The refold solution was then left for 48 h at 4°C with stirring before being concentrated to a volume of 10 mL at 15,000 x *g* in large protein concentrators (Nalgene, MWCO 10 kDa); during concentration the sample was also part buffer-exchanged by diluting with 200 mM Tris-HCl, pH 8.0, 200 mM NaCl. A sample was taken for SDS-PAGE prior to loading onto a 16/60 HiLoad Superdex 75 or 200 gel filtration column pre-equilibrated with 200 mM Tris-HCl, pH 8.0, 200 mM NaCl on an ÄKTAexpress (GE healthcare) for complete buffer-exchange and purification.

SDS-PAGE gels showing the refold results for truncations #6, #25, #31 and #95 are shown in **Fig. 4.16**. A small amount of successfully refolded protein was seen for truncation #95, see elution profile and resulting SDS-PAGE gel in **Fig. 4.17**. The elution profile revealed that a significant amount of refolded #95 was in fact soluble aggregate eluting early off the column, however there was also a single homogenous peak of soluble #95 eluting at volume suggestive of #95 monomers. To check this, the gel phase distribution coefficient (K_{av}) was calculated and checked against the calibration curve for the HiLoad Superdex 200 column (**Fig. 4.18**, a plot of K_{av} versus log molecular masses for various protein standards, GE Healthcare Calibration Kit). K_{av} was calculated using the equation below and the following parameters: elution volume ($V_e = 84$ mL), void volume ($V_o = 45$ mL for blue dextran that cannot enter the pores of the column) and total bed volume ($V_t = 120$ mL):

$$K_{av} = \frac{V_e - V_o}{V_t - V_o} = \frac{(84 \text{ mL} - 45 \text{ mL})}{(120 \text{ mL} - 45 \text{ mL})} = 0.50$$

Figure 4.16. SDS-PAGE gels of denatured and refolded truncations.

(A) Samples #31, #25 and #6. **(B)** Sample #95. The molecular masses for #6, #25, #31 and #95 expressed from pHAR1119 are 32.5 kDa, 34.1 kDa, 37.3 kDa and 44.0 kDa, respectively. Arrows indicate the positions of protein bands at the correct molecular mass (#6 run lower as observed previously); D = denatured (solubilised) protein in GnHCl buffer pre-fold (stored in 100 mg aliquots); R = refolded protein concentrated to a volume of 10 mL (prior to loading onto gel filtration column); – = empty lane; Mk^L is low range SigmaMarker (Sigma) with molecular masses (kDa) for the protein ladder listed on the right of each gel.

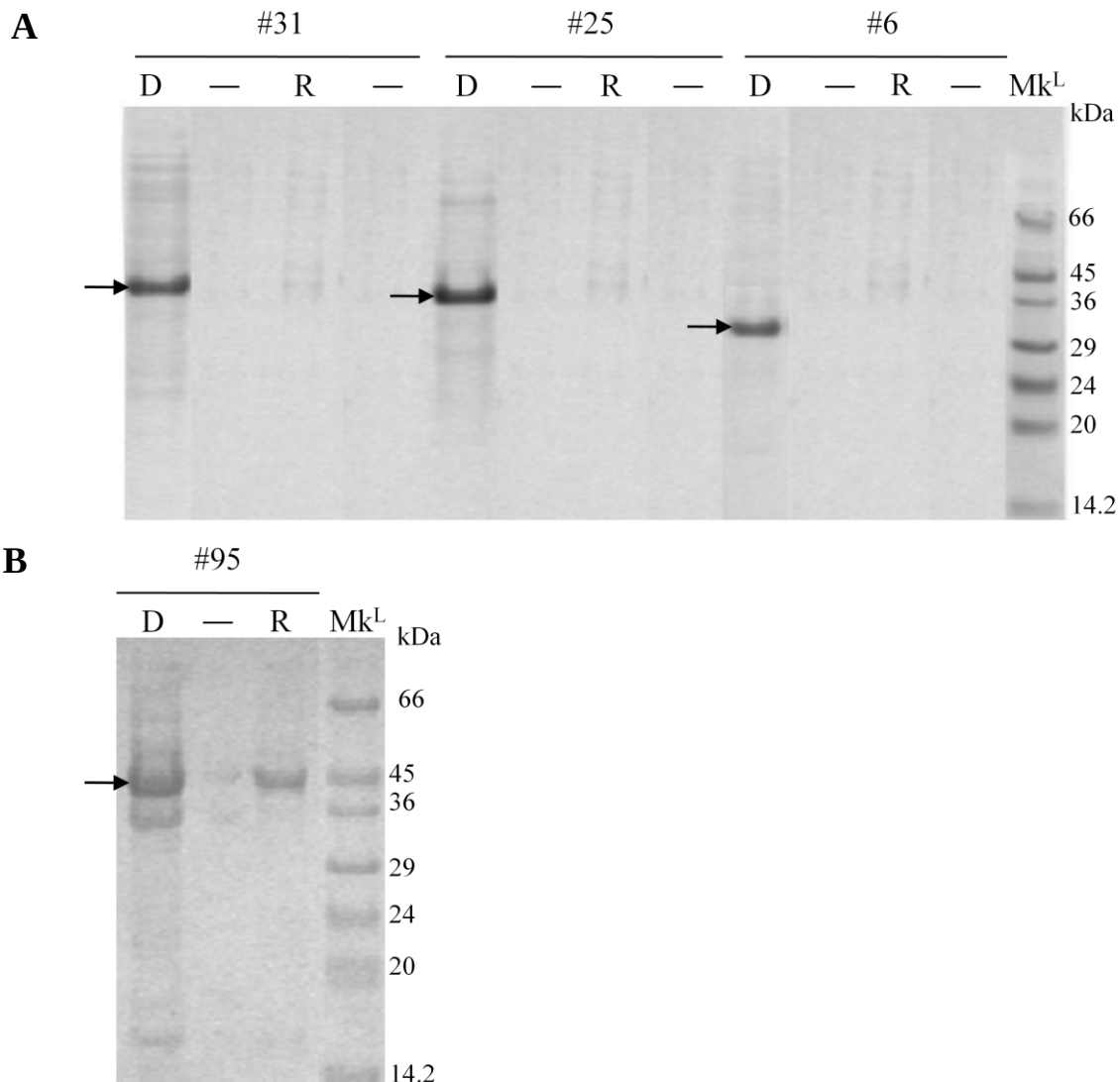
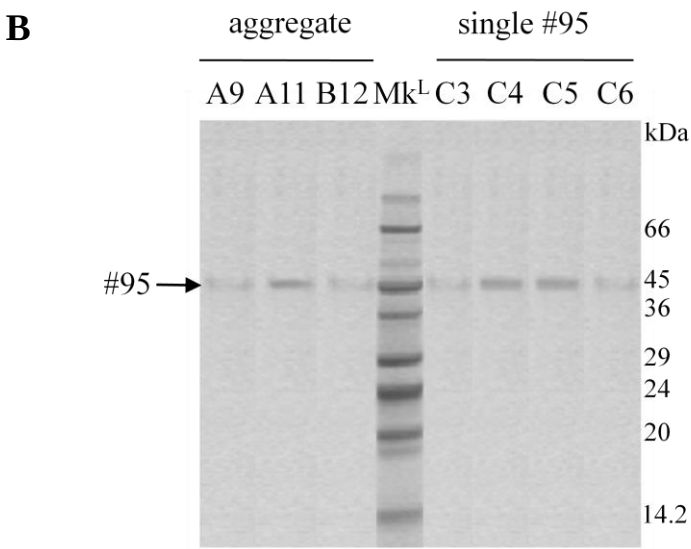
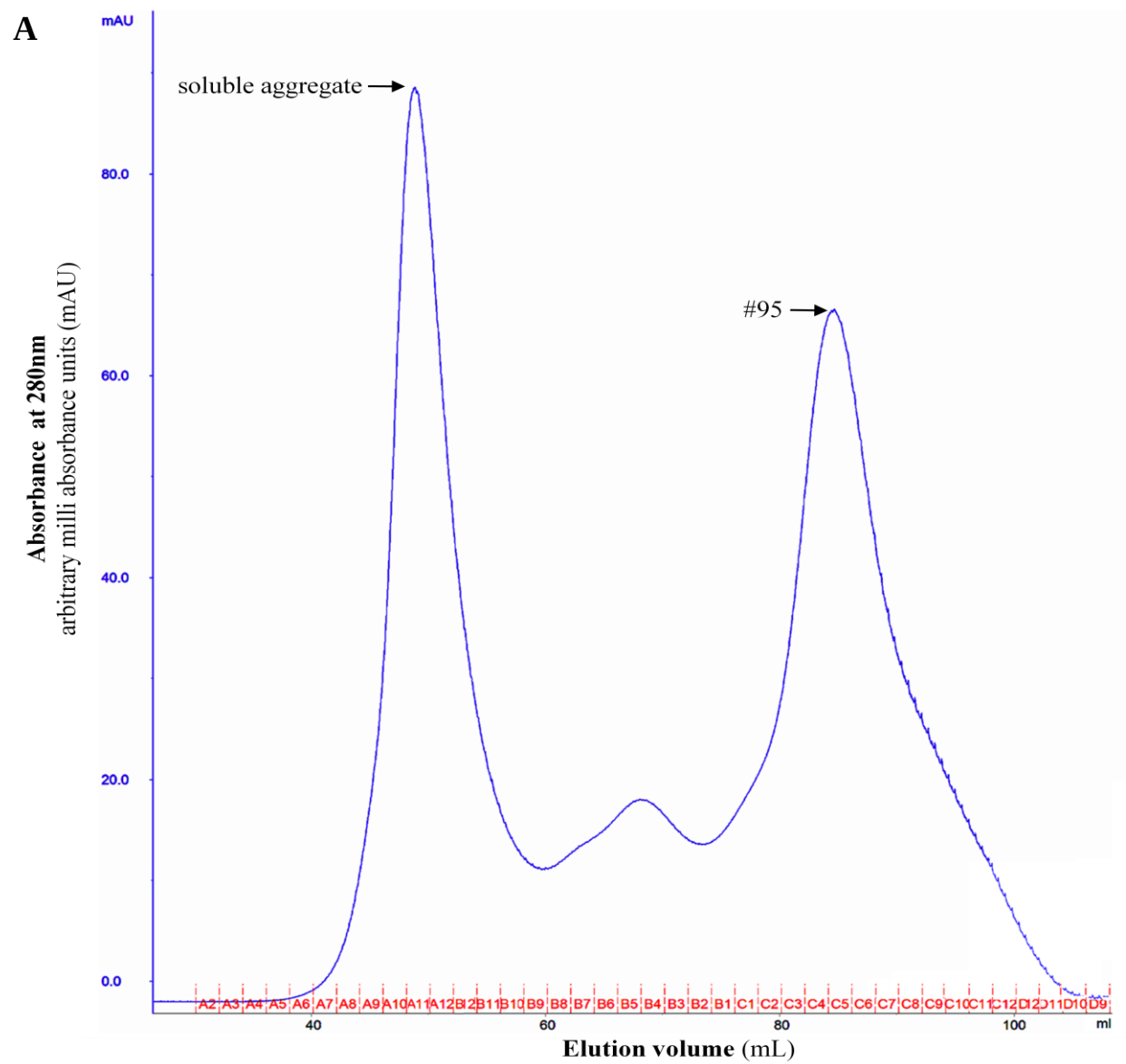


Figure 4.17 (Overleaf). Elution profile and SDS-PAGE of eluted fractions for refolded #95. (A) Elution profile (A_{280}) of proteins from gel filtration column with fractions numbered and indicated in red. Arrows indicate the position of soluble aggregate (elution volume, V_e = 50 mL) and single monomers of truncation #95 (V_e = 84 mL). **(B)** SDS-PAGE gel of eluted fractions A9, A11 and B12 spanning the soluble aggregate peak, and C3-C6 spanning the monomer peak (fractions are numbered in accordance with A); Mk^L = low range SigmaMarker (Sigma).



The K_{av} for the peak eluting at 84 mL was 0.50 which is equivalent to a molecular mass of approximately 50 kDa from the calibration curve (**Fig. 4.18**). The calculated molecular mass of truncation #95 is 44 kDa indicating that these fractions likely contained successfully refolded monomers of #95. In order to confirm this observation, a 15 μ L protein sample (20 μ M) was taken for liquid chromatography-electrospray ionisation-mass spectrometry (LC-ESI-MS); this experiment was carried out by Dr Joanne Nettleship using the quality assurance service provided by the OPPF (Nettleship *et al.*, 2008). The resulting mass spectrum showed a single peak at 44,054 Da (**Fig. 4.19**) confirming the presence of truncation #95 in the eluted fractions.

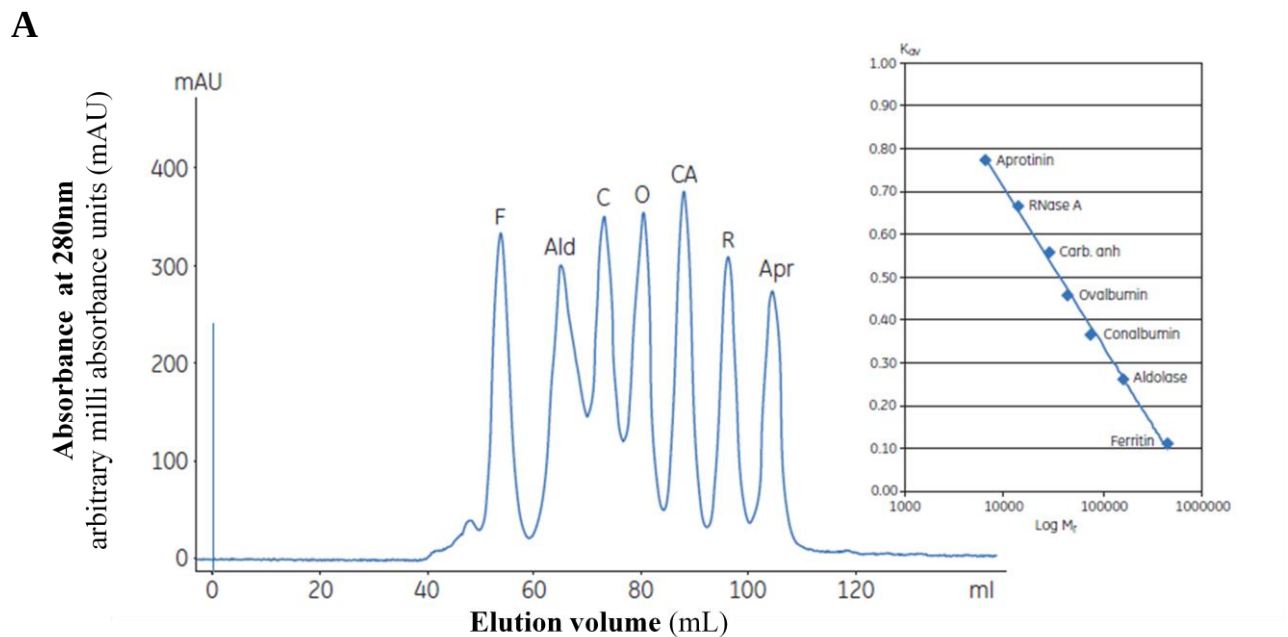
The refold efficiency was approximately 5% (5 mg successfully refolded soluble truncation #95 out of 100 mg denatured protein) which was quite poor. Unfortunately, refolding truncation #95 was not readily reproducible using this protocol and in some cases did not yield any refolded protein (not even soluble aggregate). In addition, truncation #95 began to precipitate out of the gel filtration buffer (200 mM Tris-HCl, pH 8.0, 200 mM NaCl) after storing for 48 h at 4°C, suggesting it was not stably soluble and easily aggregated. In an attempt to increase the amount of refolded #95 the following changes were made:

- (i) redox pair changed to 6.5 mM cysteamine and 3.7 mM cystamine;
- (ii) refolding at lower protein concentration (0.05 mg/mL protein in refold solution;
- (iii) changing the ionic strength (100 mM NaCl – 500 mM NaCl, and no salt).

None of these changes improved the yield or reproducibility of refolded #95 and unfortunately insufficient protein was obtained for crystallisation trials. These results however do provide direct evidence that truncation #95 ($^{433}\text{SSVEL-QGLVA}^{777}$), covering most of the putative RdRp domain (**Fig. 4.6.** and **Fig. 4.11.**) and containing hot spot sequence $^{433}\text{SSVEL}$ (**Fig. 4.6** and section 4.5.2.), can be successfully refolded from inclusion bodies in *E. coli* albeit at a low yield.

Figure 4.18. Calibration curve (Kav versus log molecular masses).

A calibration kit (GE Healthcare) was used to generate an elution profile for 7 protein standards on the HiLoad 16/60 Superdex 200 gel filtration column. The elution volumes (V_e) were used to calculate the gel phase distribution coefficient (K_{av} , following the equation described in section 4.6.2.4) for the 7 protein standards of known molecular mass. K_{av} was then plotted against the logarithm of molecular masses and a calibration line drawn. **(A)** Elution profile and K_{av} vs. log molecular mass plot; figure adapted from calibration kit. **(B)** Table listing protein standards used from the calibration kit and their molecular masses (kDa); acronyms used in (A) to identify the different protein peaks are shown in brackets in the table.

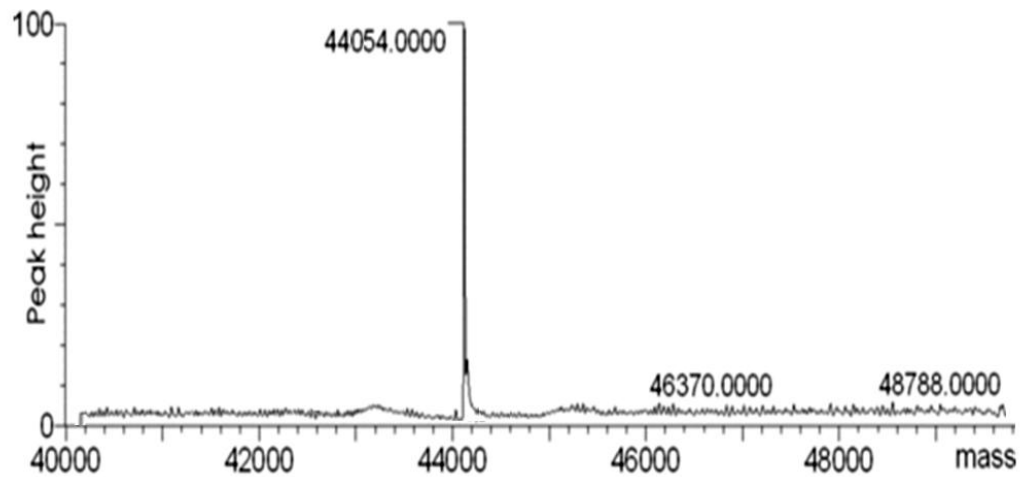


B

Protein standards	Molecular mass (kDa)
Aprotinin (Apr)	6.5
RNase A (R)	13.7
Carbonic anhydrase (CA)	29.0
Ovalbumin (O)	43.0
Conalbumin (C)	75.0
Aldolase (Ald)	158.0
Ferritin (F)	440.0

Figure 4.19. Mass Spectrum (LS-ESI-MS) of truncation #95.

The mass spectrum indicates a clear m/z peak for truncation #95 at 44,054 Da. This is consistent with the calculated mass of 44,000 Da (44 kDa) using the ExPASy ProtParam tool (Gasteiger *et al.*, 2003) with the #95 protein sequence including His₆-TEV and BAP tags provided by the pHAR1119 vector .



4.7. SARS NSP-12 – Analysis of Results

Sequencing of the gene truncations produced by ESPRIT showed a pattern of strong clustering across the length of the NSP-12 gene particularly in the C-terminal two thirds of the gene containing the putative RdRp region. Unfortunately, gene truncations in the largest clusters did not show signs of soluble expression in *E. coli*. In fact at small scale, clustered truncations consistently show up as a protein band of approximately 6.5 kDa, identified from the denaturing purification of insoluble pellets (**Fig. 4.9.**), suggesting that they had been degraded by bacterial proteases. Overall, the small scale expression results from the 95 truncations expressed from pHAR1119 vector were somewhat disappointing with few soluble hits identified that could be over-expressed and purified from *E. coli*. However, a closer look at truncations that did show signs of expression in *E. coli* identified 10 truncations (#6, #10, #16, #21, #27, #47, #68, #78, #86 and #88) that contained putative RdRp motif start or end sequences (see **Fig. 4.10.** and **Fig. 4.6.**); this suggested that the ESPRIT experiment had successfully ‘trapped’ expressable domains with defined boundaries that are equivalent to sequence conserved regions essential for RdRp activity. Five of these truncations (#6, #16, #21, #86 and #88) showed significant soluble expression at small scale supported by Western blots, whereas others ended up in inclusion bodies at the correct molecular mass (**Table 4.1**).

Twelve truncations in total were chosen for large scale expression screening, including top soluble hits with NSP-12 RdRp domain boundaries and clustered truncations containing ‘hot spot’ sequences (**Fig. 4.6**) that only showed insoluble expression. At large scale, gene truncations #6 and #95 gave the most promising results. Truncation #6 showed evidence for soluble expression when produced in *E. coli* C41 (DE3) cells (for which ~1mg of protein was obtained after purification) and truncation #95 could be over-expressed in the inclusion

bodies of *E. coli* BL21 (DE3) RIL cells and then successfully refolded by rapid dilution (for which ~5mg of protein was obtained after gel filtration). In both instances a significant amount of soluble aggregate was present (eluting in or near the void volume of the gel filtration column). For truncation #95 however, these soluble aggregates were separated from the population of #95 monomers during the gel filtration run. Unfortunately, refolding of truncation #95 was not always reproducible and the protein had a tendency to precipitate after purification. Thus, insufficient yield was obtained for crystallisation.

The main limiting factor of this work is protein insolubility. Many gene truncations are expressible in *E. coli*, but end up in inclusion bodies at both small and large scale. In fact the insoluble expression levels of different gene truncations vary quite drastically, even for closely related truncations. This suggests that the choice of domain boundaries is critical for NSP-12 expression and implies that it is not readily suitable for structure determination by standard methods. However, it is important to consider the limitations of the experiment. SARS NSP-12 is a 2796 bp gene and in sub-library 2 we selected for gene truncations from 1000 – 2796 nucleotides, a window of 1796. There could therefore be 1,612,808 possible DNA molecules produced [2796 bp gene – [(upper 2796 cut-off + lower 1000 cut-off)/2] x 1796 nucleotide window]. Considering that only 1 in 9 clones will be in-frame with both His₆- and BAP-tags, there will only be 179,200 possible complete open reading frames (ORFs). Assuming all 27,648 colonies (72 x 384-well plates) picked and gridded in the experiment represent good clones (i.e. 100% contain a correct gene truncation insert, as observed for the 24 colonies PCR-screened and sequenced for sub-library 2), only 1 in 9 of these will be in-frame with both tags; thus, only 3,072 complete ORFs are sampled. The total diversity screened in the experiment is therefore 3,072 complete ORFs out of the 179,200

ORFs produced, which is 1.7%. 98.3% of gene truncations in complete ORFs that could give rise to isolated domains that express solubly in *E. coli*, are not sampled in the experiment.

For NSP-12, the assumption that the BAP tag is only accessible for biotinylation by BirA if the protein is soluble, is not fully supported by the expression screens detailed herein. The low number of soluble hits at small scale likely reflects the poor readout of the ESPRIT biotinylation method for the solubility of the NSP-12 truncations expressed in *E. coli*. However, success using this ESPRIT method has been reported in the literature [e.g. crystal structure of the C-terminal domain of influenza virus polymerase PB2 subunit (Tarendeau *et al.*, 2007)] which suggests that the reporter system is sensitive to the protein target of interest. More surprisingly, the data herein suggest that the BAP tag provided by the pHAR1119 vector is required for soluble expression. Loss of expression is almost always seen after re-cloning into vectors lacking the BAP tag (e.g. pOPINF), suggesting that it can act as a solubility enhancer.

In summary, using the ESPRIT method it has been possible to identify for the first time in a systematic fashion, conditions under which the SARS polymerase NSP-12 protein can be solubly expressed at small scale in *E. coli*. Large-scale expression is currently problematic, and further work will be necessary to define experimental conditions under which the protein can be produced and (hopefully) crystallised. Nonetheless, this work may serve as a key first step in structurally characterising the NSP-12 polymerase protein (a key drug target of the coronavirus family), especially in terms of domain boundaries.

Chapter 5

Conclusions and Future Research

5.1. ϕ 6-RdRp – Conclusions

The structural and biochemical studies on ϕ 6 RdRp described in this thesis build on our current understanding of ϕ 6 RdRp-catalysed RNA polymerisation. We are fortunate that the $P2_1$ crystal form of ϕ 6 RdRp has allowed us to capture different complexes of the polymerase within the same asymmetric unit; this is due to differences in the crystal contacts of the three copies of ϕ 6 RdRp, varying the degree of flexibility for ligand binding and accommodating structural changes upon *in crystallo* polymerisation. The different complexes described herein shine some light on the molecular details of template and NTP entry during preinitiation, transition from initiation to elongation with displacement of the CTD, and the role of the non-catalytic Mn^{2+} during these different stages. A summary model is provided in **Fig. 3.23**.

Thermal denaturation experiments (described in section 3.1) demonstrated that both WT and E634Q mutant RdRps have an enhanced molecular flexibility when Mn^{2+} is bound at the non-catalytic ion site and an increased structural stability when out-competed by Mg^{2+} at a concentration of 10 mM. This observation suggested a possible link between the structural destabilisation required for elongation (i.e. displacement of the CTD and shifting down of the template) and bound Mn^{2+} at the non-catalytic ion site.

A number of attempts were then made to capture a complex of $\phi 6$ RdRp with an elongated dsRNA product (third nucleotide added to daughter template strand) bound at the active site and CTD displaced to its new position. However, when the elongation reaction was run *in crystallo* the crystals were observed to very quickly crack and disintegrate, preventing the collection of useful data; this was likely due to molecular movements of the CTD upon elongation. In an attempt to reduce the amount of crystal handling and speed up the mounting of crystals for data collection, 20% (v/v) glycerol cryoprotectant was added to the crystal drop as part of the soak solution containing polymerase substrates, rather than transferring the crystal to a separate cryoprotectant drop before mounting. This approach allowed direct mounting from the crystal drop after initiating *in crystallo* polymerisation by soaking in substrates. Although we did not obtain an elongation complex using this method, the resulting E634Q-hairpin-AMPPCP-Mg²⁺ complexes have captured the following molecular movements within the polymerase which have not been observed previously: opening of double-stranded hairpin oligonucleotide template; shifting down of template during the transition to elongation with concomitant displacement of the CTD; and transient weakening of the high-affinity non-catalytic ion site during transition resulting in the release of bound ‘destabilising’ Mn²⁺.

In addition to providing direct evidence for CTD displacement and template exit, the electron density maps for the different E634Q-hairpin-AMPPCP-Mg²⁺ complexes also demonstrated that incoming NTPs can bring Mg²⁺ ions with them to the active site for use in catalysis. These structures allowed us to identify a negatively charged Mg²⁺ stabilising pocket that sits on the opposite side of the NTP tunnel to the positively charged interrogation site, and is formed by D324, D327 and D329, and the backbone carbonyl groups of D324, V325 and S326. These structures indicate that Mg²⁺ ions (drawn into the polymerase by NTPs) are

stabilised about this pocket during NTP entry before being drawn to their catalytic position with the NTP movement to a base pairing position with bound template. Furthermore, crystal structures of WT-Tri4T-ATP-Mg²⁺, WT-5'-AATCT-3'-ATP-GTP-Mg²⁺ and E634Q-5'-UUGCCC-3'-GTP-Mg²⁺ revealed two NTPs stabilised at the active site in a *de novo* initiation-like position in the absence of any bound template. In these structures, template entry was hindered by the replacement of the non-catalytic ion site Mn²⁺ with Mg²⁺ (by soaking with 10 mM MgCl₂). These structures showed how Mg²⁺ can adopt a number of geometries about the non-catalytic ion site ('downward', 'centred' or 'upward') during pre-initiation so as to stabilise the negative charge of the bound NTPs in the absence of bound template.

Opening of the double-stranded hairpin template was evident from the WT-Tri4T-Mg²⁺ structure which demonstrated how interactions with the molecular surface of the RdRp can facilitate denaturation of oligonucleotide secondary structures before the template enters into the active site. In addition, crystal structures of WT-5'-AATCT-3'-GTP-Mg²⁺ and WT-5'-AATCT-3'-ATP-GTP-Mg²⁺ revealed a positively charged groove on the outer surface of ϕ 6 RdRp (including key residues R30 and K541) that directs the incoming oligonucleotide template into the polymerase and towards the active site. These data are in close agreement with previous mutagenesis data on R30 and K541 in template binding (Sarin *et al.*, 2009).

Finally, considering that our data reveal that Mg²⁺ can replace the non-catalytic Mn²⁺ in ϕ 6 RdRp when crystals are soaked with 10 mM MgCl₂, we carried out an analysis on the crystallization conditions of other published viral RdRp structures. This analysis revealed that high concentrations of different Mg²⁺ salts (up to 300 mM) were present during crystallization of many of these viral RdRps, questioning the authenticity of the Mg²⁺ ion at

the equivalent non-catalytic ion site in these previously published structures (**Table 3.2**). Our research reveals that Mn^{2+} bound at this site in $\phi 6$ RdRp plays an important role in template binding and controlling the switch to elongation. In addition we demonstrate that although Mg^{2+} can occupy this site, it cannot functionally replace the natural ion (Mn^{2+}). It is likely that the function of the non-catalytic ion site in controlling the switch from initiation to elongation is broadly conserved in viral RdRps.

5.2. $\phi 6$ -RdRp – Future Research

One remaining question for the $\phi 6$ -catalysed polymerisation reaction is whether initiation of replication follows the same mechanism of initiation reported for transcription (Butcher *et al.*, 2001). This experiment was carried out within the scope of this research using a DNA oligonucleotide mimic of (+)-sense ssRNA (see for example structures WT-5'-AATCT-3'-GTP- Mg^{2+} and WT-5'-AATCT-3'-ATP-GTP- Mg^{2+}), however the divalent cation conditions did not permit template entry into the active site. Instead these structures provided detailed information about the direction of template entry on the surface of the polymerase and key residues that are involved in template binding and threading down of template into the active site. In order to characterise the initiation of replication, a future experiment should aim to run the initiation reaction *in crystallo* with the 3'-end of (+)-sense genomic RNA (e.g. 5'...UCU-3') and WT $\phi 6$ RdRp under the same conditions used to capture the initiation of the transcription complex (Butcher *et al.*, 2001).

In addition to its polymerase activity, $\phi 6$ RdRp has also been shown to have terminal nucleotidyl transferase (TNTase) activity, catalysing the addition of nucleotides to the 3'-end of either ssRNA or dsRNA (preferred by $\phi 6$ RdRp). RNA polymerases of ssRNA viruses,

such as HCV, BVDV, GB virus-B, NV, SV and PV also possess TNTase activity on ssRNA substrates (Fullerton *et al.*, 2007; Neufeld *et al.*, 1994; Ranjith-Kumar *et al.*, 2001; Ranjith-Kumar *et al.*, 2003; Rohayem *et al.*, 2006; Zhong *et al.*, 1998). It has been suggested that this activity is designed to repair the 3'-ends of viral genomes that may have been subjected to degradation by cellular exonucleases (Ranjith-Kumar *et al.*, 2001) although a precise biological function in these viruses remains elusive.

In $\phi 6$ RdRp a model for TNTase activity, exploiting the two metal ion catalysis mechanism used for polymerisation, was suggested based on biochemical observations (Poranen *et al.*, 2008a). In contrast to the template-dependent polymerase activity, which requires opening of the dsRNA template and feeding of ssRNA into the template tunnel, the TNTase reaction apparently depends on opening of the compact apoenzyme structure at the CTD (similar to the displacement observed during elongation) allowing direct access of the dsRNA to the active site. The ability to accept and replicate artificial circular templates also indicates that the $\phi 6$ RdRp apoenzyme may adopt different conformations which allow the polymerase to enclose circular templates (Ranjith-Kumar and Kao, 2006).

Structural flexibility seems to be an intrinsic property of $\phi 6$ RdRp and Mn^{2+} bound at the non-catalytic ion site in the purified polymerase is likely to play a key role in increasing structural fluidity required for CTD repositioning; see thermal denaturation experiments (section 3.1). In line with this, Poranen *et al.*, (2008a) reported that the TNTase activity of the E491Q mutant polymerase was less than 10% of WT $\phi 6$ RdRp, which reflects the reduced affinity of the E491Q mutant for Mn^{2+} at the non-catalytic ion site (resulting increased structural rigidity) and absence of Mn^{2+} bound in the purified E491Q apoenzyme despite its presence in the crystallization buffer (Poranen *et al.*, 2008a). Future structural work should

attempt to validate the current model for terminal nucleotidyl addition (Poranen *et al.*, 2008a) and identify whether Mn^{2+} bound at the non-catalytic ion site is required for activity.

Structural and biochemical studies on $\phi 6$ RdRp described herein provide a detailed understanding of the molecular details of *de novo* polymerisation (from preinitiation through to elongation), and serve as a paradigm for other structurally related polymerases, such as HCV and BVDV polymerases. For polymerases that initiate *de novo* RNA synthesis, understanding the molecular movements of the CTD and effect of different divalent cations throughout the polymerisation cycle is still an active area of research. It is likely that future research into the role of the CTD in initiation of replication and TNTase activity will also provide further molecular insight that is applicable to other viral RdRps.

5.3. SARS-CoV NSP-12 – Conclusions

DNA fragmentation-based combinatorial approaches to soluble protein expression are increasingly being used for challenging proteins (such as SARS-CoV NSP-12) that are not readily expressible using conventional methods. One such approach, known as ESPRIT (Expression of Soluble Protein by Random Incremental Truncation), was used in this research to identify (or trap) protein truncations of SARS-CoV NSP-12 that could be expressed solubly in *E. coli* and therefore be suitable for crystallisation and structure determination by X-ray crystallography.

The top 95 hits identified by the ESPRIT method were sequenced and expressed at small-scale using high-throughput protein expression protocols at the Oxford Protein Production Facility (Berrow *et al.*, 2007; Berrow *et al.*, 2006). The small scale expression results from

the 95 truncations expressed from the pHAR1119 vector (which adds a biotin acceptor peptide to the expressed proteins) were somewhat disappointing with fewer than expected soluble hits identified. In an attempt to yield more small scale soluble hits from the ESPRIT top 95, a number of changes were made to the small scale expression protocol (for example, changes in expression temperature and IPTG concentration); this did not increase the number of soluble hits identified.

Aligning the top 95 hits from the ESPRIT experiment along the length of NSP-12 revealed six main clusters (1-6) of gene truncations in particular ‘hot spot’ regions. Disappointingly very few of the gene truncations starting and ending at the same point (forming clusters along the length of NSP-12), gave rise to significant soluble protein expression. A closer look at the truncations that did show signs of expression in *E. coli* identified 10 truncations (#6, #10, #16, #21, #27, #47, #68, #78, #86 and #88) that contained putative RdRp motif start or end sequences, suggesting that the ESPRIT experiment had successfully ‘trapped’ expressable domains with defined boundaries that are equivalent to sequence conserved regions essential for RdRp activity. Five of these truncations (#6, #16, #21, #86 and #88) showed significant soluble expression at small scale supported by Western blots; the others hits were over-expressed but ended up in inclusion bodies. Truncations #16 and #21 were grouped in hot-spot cluster 5 (truncations ending with ⁶⁷⁸GGTSS; see section 4.5.2). Interestingly, ⁶⁷⁸GGTSS is the precise starting sequence of putative NSP-12 RdRp motif B, providing a further level of confidence that the ESPRIT method can identify specific domain boundaries.

Large-scale protein production was carried out for 12 of the protein truncations identified at small scale following the high-throughput protocols of the Oxford Protein Production Facility (Berrow *et al.*, 2007; Berrow *et al.*, 2006). The results were somewhat disappointing with

only gene truncation #6 (⁶⁹⁶ICQAV-HTVLQ⁹³²) showing very weak soluble expression and most ending up in inclusion bodies. In order to improve soluble expression of the twelve truncations at large scale, a number of rescue strategies were attempted including: expression of truncations in different *E. coli* strains, and re-cloning of gene truncations into different pOPIN expression vectors to remove the biotin acceptor peptide (BAP) tag and replace it with either a His₆-, SUMO- or GST-tag; it is worth noting that RdRp-GST fusions have been reported previously for example with NV and RHDV polymerases, (Ng *et al.*, 2004; Vazquez *et al.*, 1998). In addition, a synthetic truncation (⁴³³SSVEL-ILGAG⁸⁴¹), covering the entire putative RdRp region and containing hot-spot cluster 4 and cluster 6 sequences (see section 4.5.2) was designed and expression-tested at large scale. These approaches had limited success in producing soluble protein at large scale. As a final experiment, refolding of over-expressed truncations #6, #25, #31 and #95 from inclusion bodies was attempted. Truncation #95 (⁴³³SSVEL-QGLVA⁷⁷⁷), covering most of the putative RdRp region and containing hot-spot cluster 4 starting sequence ⁴³³SSVEL (see section 4.5.2), was the only truncation that showed signs of successful refolding, albeit at a low efficiency (5%). Unfortunately refolding of #95 was not readily reproducible and sufficient yield was not obtained for crystallization experiments. However, this experiment demonstrated for the first time refolding of an NSP-12 truncation at large scale which should be explored further by future research.

Overall, the ESPRIT method described in this thesis has allowed us to identify for the first time in a systematic fashion, conditions under which the SARS polymerase NSP-12 protein can be solubly expressed at small scale in *E. coli*. Large-scale expression is currently problematic, and further work will be necessary to define experimental conditions under which the protein can be produced and (hopefully) crystallised. Nonetheless, this work

serves as a key first step in structurally characterising the NSP-12 polymerase protein (a key drug target of the coronavirus family), especially in terms of domain boundaries.

5.4. SARS-CoV NSP-12 – Future Research

In this research the ESPRIT method was used to identify soluble domains of SARS-CoV NSP-12 that could be expressed in *E. coli*. Soluble fragments of NSP-12 were identified at small scale but translating to large scale expression was problematic. As an extension to this work, insect cell (baculovirus) expression or other eukaryotic expression systems should be tried using the soluble targets from the small scale *E. coli* experiments. A good starting point would be with the five protein truncations (#6, #16, #21, #86 and #88) containing the putative RdRp motifs that showed significant soluble expression in *E. coli* at small-scale (but did not yield soluble protein at large scale), and protein truncation #95 that showed evidence of refolding at large scale. The greater similarity of eukaryotic expression systems (in terms of protein folding machinery, cytoplasmic environment and post-translational modifications) to the host eukaryotic cells infected by SARS-CoV, are more likely to encourage the ‘trapped’ NSP-12 domains to fold correctly and not aggregate into inclusion bodies when scaled up in large scale experiments.

A further project could look at the potential of running a different high-throughput DNA-fragmentation approach on the SARS-CoV NSP-12 gene that provides a more reliable readout for solubility when screening constructs expressed in a colony array. We demonstrate in this research that biotinylation of protein truncations can act as a solubility enhancer. This has been reported for similar library-based expression and screening approaches that use tags such as Green Fluorescent Protein (GFP), (Waldo, 2003). However, in attempt to counter the passenger solubilisation properties of the GFP tag, Waldo and co-workers have further

developed their GFP tag methodology for screening a library of constructs. The new methodology is based on a two plasmid split-GFP complementation system (Cabantous *et al.*, 2005a; Cabantous *et al.*, 2005b; Cabantous and Waldo, 2006) whereby GFP is divided into an 11-residue fragment used to tag the expressed proteins first, followed by the separate induction of the remaining 227-residues of GFP. Only when the 11-residue tag of the expressed target is available to combine with the remainder of the GFP does fluorophore-maturation take place. This approach would provide an alternative means to carry out high-throughput assessment of protein expression in either colonies or small-scale liquid culture formats, providing of course that the 11-residue fragment of the GFP does not perturb the correct folding of the target protein. It would be interesting to see whether Waldo's methodology could screen more accurately for fully soluble protein domains of NSP-12 and whether identified targets contained similar domain boundaries to those in the ESPRIT method that contain RdRp sequence motifs.

An alternative approach would be to co-express full-length SARS-CoV NSP-12 (or one of the previously identified soluble domains at small scale from *E. coli*) with full-length NSP-13 (HEL1, helicase protein). Yeast-2-hybrid (Y2H) assays and follow-up co-immunoprecipitation experiments were carried out by von Brunn *et al.*, (2007) in order to determine intraviral protein-protein interactions for SARS-CoV. The Y2H bait/prey assay indicated that NSP-12 interacts in both directions with NSP-13 (HEL1), NSP-8 (primer synthesis and processivity factor in combination with NSP-7) and ORF-9b (accessory protein involved in virion assembly); these results were supported by co-immunoprecipitation experiments (von Brunn *et al.*, 2007). Structures of the NSP-7/NSP-8 complex and ORF-9b are already available (Meier *et al.*, 2006; Zhai *et al.*, 2005), however, like with NSP-12, expression and crystallization of the HEL1 protein has been hindered by poor solubility and

low yield. One hypothesis would be that soluble expression of NSP-12 requires assistance and/or binding of co-expressed HEL1 to form a complex that masks hydrophobic surface-exposed protrusions on both NSP-12 and HEL1. To test this hypothesis, the proteins should be co-expressed in either *E. coli* or a eukaryotic system. To purify the potential complex by affinity chromatography, a His₆-tag could be added to NSP-12 and a Steptavidin tag (Strep-tag) to HEL1. Purification could be carried out first by Ni²⁺-affinity chromatography to purify proteins containing a His₆-tag and then passed through a Strep-Tactin affinity column to purify Strep-tagged proteins; this will allow purification of the individual proteins (NSP-12 and HEL1 separately) if the complex has not formed. If no soluble expression is observed a denaturing purification on the insoluble pellet should be performed in order to assess whether the complex or individual proteins were successfully expressed at all.

Finally, our refolding experiments demonstrate that one of the top 95 ESPRIT hits (truncation #95) could be successfully refolded using a rapid dilution protocol. A similar refold approach was used for example to obtain sufficient soluble SARS-CoV ORF7a accessory protein for crystallisation and subsequent structure determination (Nelson *et al.*, 2005). Future research should explore different refold strategies for truncation #95 and other RdRp truncations discussed herein, in order to obtain sufficient protein for crystallization trials and hopefully subsequent structure determination.

References

- Ackermann, M., and Padmanabhan, R. (2001). De novo synthesis of RNA by the dengue virus RNA-dependent RNA polymerase exhibits temperature dependence at the initiation but not elongation phase. *J Biol Chem* 276, 39926-39937.
- Adams, P.D., Grosse-Kunstleve, R.W., Hung, L.W., Ioerger, T.R., McCoy, A.J., Moriarty, N.W., Read, R.J., Sacchettini, J.C., Sauter, N.K., and Terwilliger, T.C. (2002). PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr D Biol Crystallogr* 58, 1948-1954.
- Ago, H., Adachi, T., Yoshida, A., Yamamoto, M., Habuka, N., Yatsunami, K., and Miyano, M. (1999). Crystal structure of the RNA-dependent RNA polymerase of hepatitis C virus. *Structure* 7, 1417-1426.
- Alaoui-Lsmaili, M.H., Hamel, M., L'Heureux, L., Nicolas, O., Bilimoria, D., Labonte, P., Mounir, S., and Rando, R.F. (2000). The hepatitis C virus NS5B RNA-dependent RNA polymerase activity and susceptibility to inhibitors is modulated by metal cations. *J Hum Virol* 3, 306-316.
- Almeida, M.S., Johnson, M.A., and Wuthrich, K. (2006). NMR assignment of the SARS-CoV protein nsp1. *J. Biomol. NMR In Press*.
- Alzari, P.M., Berglund, H., Berrow, N.S., Blagova, E., Busso, D., Cambillau, C., Campanacci, V., Christodoulou, E., Eiler, S., Fogg, M.J., *et al.* (2006). Implementation of semi-automated cloning and prokaryotic expression screening: the impact of SPINE. *Acta Crystallogr D Biol Crystallogr* 62, 1103-1113.
- Anand K, Palm GJ, Mesters JR, Siddell SG, Ziebuhr J, and Hilgenfeld R (2002). Structure of coronavirus main proteinase reveals combination of a chymotrypsin fold with an extra alpha-helical domain. *EMBO Journal* 21, 3213-3224.
- Appleby, T.C., Luecke, H., Shim, J.H., Wu, J.Z., Cheney, I.W., Zhong, W., Vogeley, L., Hong, Z., and Yao, N. (2005). Crystal structure of complete rhinovirus RNA polymerase suggests front loading of protein primer. *J Virol* 79, 277-288.
- Arnold, J.J., Ghosh, S.K., and Cameron, C.E. (1999). Poliovirus RNA-dependent RNA polymerase (3D(pol)). Divalent cation modulation of primer, template, and nucleotide selection. *J Biol Chem* 274, 37060-37069.
- Artsimovitch, I., Patlan, V., Sekine, S., Vassilyeva, M.N., Hosaka, T., Ochi, K., Yokoyama, S., and Vassilyev, D.G. (2004). Structural basis for transcription regulation by alarmon ppGpp. *Cell* 117, 299-310.
- Bartlam, M., Yang, H., and Rao, Z. (2005). Structural insights into SARS coronavirus proteins. *Curr Opin Struct Biol* 15, 664-672.

- Beckett, D., Kovaleva, E., and Schatz, P.J. (1999). A minimal peptide substrate in biotin holoenzyme synthetase-catalyzed biotinylation. *Protein Sci* 8, 921-929.
- Behrens, S.E., Tomei, L., and De Francesco, R. (1996). Identification and properties of the RNA-dependent RNA polymerase of hepatitis C virus. *EMBO J* 15, 12-22.
- Berrow, N.S., Bussow, K., Coutard, B., Diprose, J., Ekberg, M., Folkers, G.E., Levy, N., Lieu, V., Owens, R.J., Peleg, Y., *et al.* (2006). Recombinant protein expression and solubility screening in *Escherichia coli*: a comparative study. *Acta Crystallogr D Biol Crystallogr* 62, 1218-1226.
- Berrow, N.S., Alderton, D., Sainsbury, S., Nettleship, J., Assenberg, R., Rahman, N., Stuart, D.I., and Owens, R.J. (2007). A versatile ligation-independent cloning method suitable for high-throughput expression screening applications. *Nucleic Acids Res* 35, e45.
- Biswal, B.K., Cherney, M.M., Wang, M., Chan, L., Yannopoulos, C.G., Bilimoria, D., Nicolas, O., Bedard, J., and James, M.N. (2005). Crystal structures of the RNA-dependent RNA polymerase genotype 2a of hepatitis C virus reveal two conformations and suggest mechanisms of inhibition by non-nucleoside inhibitors. *J Biol Chem* 280, 18202-18210.
- Bressanelli, S., Tomei, L., Roussel, A., Incitti, I., Vitale, R.L., Mathieu, M., De Francesco, R., and Rey, F.A. (1999). Crystal structure of the RNA-dependent RNA polymerase of hepatitis C virus. *Proc Natl Acad Sci U S A* 96, 13034-13039.
- Bressanelli, S., Tomei, L., Rey, F.A., and De Francesco, R. (2002). Structural analysis of the hepatitis C virus RNA polymerase in complex with ribonucleotides. *J Virol* 76, 3482-3492.
- Brockway, S.M., Clay, C.T., Lu, X.T., and Denison, M.R. (2003). Characterization of the expression, intracellular localization, and replication complex association of the putative mouse hepatitis virus RNA-dependent RNA polymerase. *J Virol* 77, 10515-10527.
- Bruenn, J.A. (2003). A structural and primary sequence comparison of the viral RNA-dependent RNA polymerases. *Nucleic Acids Research* 31, 1821-1829.
- Bussow, K., Cahill, D., Nietfeld, W., Bancroft, D., Scherzinger, E., Lehrach, H., and Walter, G. (1998). A method for global protein expression and antibody screening on high-density filters of an arrayed cDNA library. *Nucleic Acids Res.* 26, 5007-5008.
- Butcher, S.J., Dokland, T., Ojala, P.M., Bamford, D.H., and Fuller, S.D. (1997). Intermediates in the assembly pathway of the double-stranded RNA virus phi6. *EMBO J* 16, 4477-4487.
- Butcher, S.J., Makeyev, E.V., Grimes, J.M., Stuart, D.I., and Bamford, D.H. (2000). Crystallization and preliminary X-ray crystallographic studies on the bacteriophage phi6 RNA-dependent RNA polymerase. *Acta Crystallogr D Biol Crystallogr* 56, 1473-1475.

- Butcher, S.J., Grimes, J.M., Makeyev, E.V., Bamford, D.H., and Stuart, D.I. (2001). A mechanism for initiating RNA-dependent RNA polymerization. *Nature* 410, 235-240.
- Cabantous, S., Pedelacq, J.D., Mark, B.L., Naranjo, C., Terwilliger, T.C., and Waldo, G.S. (2005a). Recent advances in GFP folding reporter and split-GFP solubility reporter technologies. Application to improving the folding and solubility of recalcitrant proteins from *Mycobacterium tuberculosis*. *J Struct Funct Genomics* 6, 113-119.
- Cabantous, S., Terwilliger, T.C., and Waldo, G.S. (2005b). Protein tagging and detection with engineered self-assembling fragments of green fluorescent protein. *Nat Biotechnol* 23, 102-107.
- Cabantous, S., and Waldo, G.S. (2006). In vivo and in vitro protein solubility assays using split GFP. *Nat Methods* 3, 845-854.
- Campagnola, G., Weygandt, M., Scoggin, K., and Peersen, O. (2008). Crystal structure of coxsackievirus B3 3Dpol highlights the functional importance of residue 5 in picornavirus polymerases. *J Virol* 82, 9458-9464.
- Chapman-Smith, A., and Cronan, J.E.J. (1999). The enzymatic biotinylation of proteins: a post-translational modification of exceptional specificity. *Trends Biochem. Sci.* 24, 359-363.
- Cheng, A., Zhang, W., Xie, Y., Jiang, W., Arnold, E., Sarafianos, S.G., and Ding, J. (2005). Expression, purification, and characterization of SARS coronavirus RNA polymerase. *Virology* 335, 165-176.
- Cho, M.W., Richards, O.C., Dmitrieva, T.M., Agol, V., and Ehrenfeld, E. (1993). RNA duplex unwinding activity of poliovirus RNA-dependent RNA polymerase 3Dpol. *J Virol* 67, 3010-3018.
- Choi, K.H., Groarke, J.M., Young, D.C., Kuhn, R.J., Smith, J.L., Pevear, D.C., and Rossmann, M.G. (2004). The structure of the RNA-dependent RNA polymerase from bovine viral diarrhea virus establishes the role of GTP in de novo initiation. *Proc Natl Acad Sci U S A* 101, 4425-4430.
- Cornvik, T., Dahlroth, S.L., Magnusdottir, A., Herman, M.D., Knaust, R., Ekberg, M., and Nordlund, P. (2005). Colony filtration blot: a new screening method for soluble protein expression in *Escherichia coli*. *Nat Methods* 2, 507-509.
- Cramer, P., Bushnell, D.A., Fu, J., Gnat, A.L., Maier-Davis, B., Thompson, N.E., Burgess, R.R., Edwards, A.M., David, P.R., and Kornberg, R.D. (2000). Architecture of RNA polymerase II and implications for the transcription mechanism. *Science* 288, 640-649.
- Cramer, P., Bushnell, D.A., and Kornberg, R.D. (2001). Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. *Science* 292, 1863-1876.
- Cramer, P. (2002). Multisubunit RNA polymerases. *Curr Opin Struct Biol* 12, 89-97.

- Cramer, P., Armache, K.J., Baumli, S., Benkert, S., Brueckner, F., Buchen, C., Damsma, G.E., Dengl, S., Geiger, S.R., Jasiak, A.J., *et al.* (2008). Structure of eukaryotic RNA polymerases. *Annu Rev Biophys* 37, 337-352.
- Davies, J.F., 2nd, Almassy, R.J., Hostomska, Z., Ferre, R.A., and Hostomsky, Z. (1994). 2.3 A crystal structure of the catalytic domain of DNA polymerase beta. *Cell* 76, 1123-1133.
- Day, L.A., and Mindich, L. (1980). The molecular weight of bacteriophage phi 6 and its nucleocapsid. *Virology* 103, 376-385.
- de Graaff, M., Houwing, C.J., Lukacs, N., and Jaspars, E.M. (1995). RNA duplex unwinding activity of alfalfa mosaic virus RNA-dependent RNA polymerase. *FEBS Lett* 371, 219-222.
- de Haas, F., Paatero, A.O., Mindich, L., Bamford, D.H., and Fuller, S.D. (1999). A symmetry mismatch at the site of RNA packaging in the polymerase complex of dsRNA bacteriophage phi6. *J Mol Biol* 294, 357-372.
- Delano, W.L. (2002). The PyMOL Molecular Graphics System. (Delano Scientific, Palo Alto, CA, USA).
- Delarue, M., Poch, O., Tordo, N., Moras, D., and Argos, P. (1990). An attempt to unify the structure of polymerases. *Protein Eng* 3, 461-467.
- Doublie, S., Tabor, S., Long, A.M., Richardson, C.C., and Ellenberger, T. (1998). Crystal structure of a bacteriophage T7 DNA replication complex at 2.2 Å resolution. *Nature* 391, 251-258.
- Drosten C, Gunther S, Preiser W, van der Werf S, Brodt HR, Becker S, Rabenau H, Panning M, Kolesnikova L, Fouchier RA, *et al.* (2003). Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *New England Journal of Medicine* 348, 1967-1976.
- Dumon-Seignovert, L., Cariot, G., and Vuillard, L. (2004). The toxicity of recombinant proteins in *Escherichia coli*: a comparison of overexpression in BL21(DE3), C41(DE3), and C43(DE3). *Protein Expr Purif* 37, 203-206.
- Ebright, R.H. (2000). RNA polymerase: structural similarities between bacterial RNA polymerase and eukaryotic RNA polymerase II. *J Mol Biol* 304, 687-698.
- Emsley, P., and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* 60, 2126-2132.
- Ewen, M.E., and Revel, H.R. (1990). RNA-protein complexes responsible for replication and transcription of the double-stranded RNA bacteriophage phi 6. *Virology* 178, 509-519.
- Ferrari, E., Wright-Minogue, J., Fang, J.W., Baroudy, B.M., Lau, J.Y., and Hong, Z. (1999). Characterization of soluble hepatitis C virus RNA-dependent RNA polymerase expressed in *Escherichia coli*. *J Virol* 73, 1649-1654.

- Ferrer-Orta, C., Arias, A., Perez-Luque, R., Escarmis, C., Domingo, E., and Verdaguer, N. (2004). Structure of foot-and-mouth disease virus RNA-dependent RNA polymerase and its complex with a template-primer RNA. *J Biol Chem* 279, 47212-47221.
- Ferrer-Orta, C., Arias, A., Agudo, R., Perez-Luque, R., Escarmis, C., Domingo, E., and Verdaguer, N. (2006). The structure of a protein primer-polymerase complex in the initiation of genome replication. *EMBO J* 25, 880-888.
- Ferrer-Orta, C., Arias, A., Perez-Luque, R., Escarmis, C., Domingo, E., and Verdaguer, N. (2007). Sequential structures provide insights into the fidelity of RNA replication. *Proc Natl Acad Sci U S A* 104, 9463-9468.
- Finney, L.A., and O'Halloran, T.V. (2003). Transition metal speciation in the cell: insights from the chemistry of metal ion receptors. *Science* 300, 931-936.
- Frilander, M., and Bamford, D.H. (1995). In vitro packaging of the single-stranded RNA genomic precursors of the segmented double-stranded RNA bacteriophage phi 6: the three segments modulate each other's packaging efficiency. *J Mol Biol* 246, 418-428.
- Fullerton, S.W., Blaschke, M., Coutard, B., Gebhardt, J., Gorbalenya, A., Canard, B., Tucker, P.A., and Rohayem, J. (2007). Structural and functional characterization of sapovirus RNA-dependent RNA polymerase. *J Virol* 81, 1858-1871.
- Garriga, D., Navarro, A., Querol-Audi, J., Abaitua, F., Rodriguez, J.F., and Verdaguer, N. (2007). Activation mechanism of a noncanonical RNA-dependent RNA polymerase. *Proc Natl Acad Sci U S A* 104, 20540-20545.
- Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R.D., and Bairoch, A. (2003). ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res* 31, 3784-3788.
- Geerlof, A., Brown, J., Coutard, B., Egloff, M.P., Enguita, F.J., Fogg, M.J., Gilbert, R.J., Groves, M.R., Haouz, A., Nettleship, J.E., *et al.* (2006). The impact of protein characterization in structural proteomics. *Acta Crystallogr D Biol Crystallogr* 62, 1125-1136.
- Glusker, J.P., Katz, A.K., and Bock, C.W. (1999). Metal ions in biological systems. *Rigaku J* 16, 8-16.
- Gnatt, A.L., Cramer, P., Fu, J., Bushnell, D.A., and Kornberg, R.D. (2001). Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 Å resolution. *Science* 292, 1876-1882.
- Gosert, R., Kanjanahaluethai, A., Egger, D., Bienz, K., and Baker, S.C. (2002). RNA replication of mouse hepatitis virus takes place at double-membrane vesicles. *J Virol* 76, 3697-3708.

- Gottlieb, P., Metzger, S., Romantschuk, M., Carton, J., Strassman, J., Bamford, D.H., Kalkkinen, N., and Mindich, L. (1988). Nucleotide sequence of the middle dsRNA segment of bacteriophage phi 6: placement of the genes of membrane-associated proteins. *Virology* 163, 183-190.
- Gottlieb, P., Strassman, J., and Mindich, L. (1992). Protein P4 of the bacteriophage phi 6 procapsid has a nucleoside triphosphate-binding site with associated nucleoside triphosphate phosphohydrolase activity. *J Virol* 66, 6220-6222.
- Grimes, J.M., Burroughs, J.N., Gouet, P., Diprose, J.M., Malby, R., Zientara, S., Mertens, P.P., and Stuart, D.I. (1998). The atomic structure of the bluetongue virus core. *Nature* 395, 470-478.
- Gruez, A., Selisko, B., Roberts, M., Bricogne, G., Bussetta, C., Jabafi, I., Coutard, B., De Palma, A.M., Neyts, J., and Canard, B. (2008). The crystal structure of coxsackievirus B3 RNA-dependent RNA polymerase in complex with its protein primer VPg confirms the existence of a second VPg binding site on Picornaviridae polymerases. *J Virol* 82, 9577-9590.
- Hansen, J.L., Long, A.M., and Schultz, S.C. (1997). Structure of the RNA-dependent RNA polymerase of poliovirus. *Structure* 5, 1109-1122.
- Hart, D.J. (2004). Systematic approach for generating soluble protein variants. (UK patent office application).
- Hart, D.J., and Tarendeau, F. (2006). Combinatorial library approaches for improving soluble protein expression in *Escherichia coli*. *Acta Cryst. D* 62.
- Henikoff, S. (1984). Unidirectional digestion with exonuclease III creates targeted breakpoints for DNA sequencing. *Gene* 28, 351-359.
- Hong, Z., Cameron, C.E., Walker, M.P., Castro, C., Yao, N., Lau, J.Y., and Zhong, W. (2001). A novel mechanism to ensure terminal initiation by hepatitis C virus NS5B polymerase. *Virology* 285, 6-11.
- Joseph, J.S., Saikatendu, K.S., Subramanian, V., Neuman, B.W., Brooun, A., Griffith, M., Moy, K., Yadav, M.K., Velasquez, J., Buchmeier, M.J., *et al.* (2006). Crystal Structure of Nonstructural Protein 10 from the Severe Acute Respiratory Syndrome Coronavirus Reveals a Novel Fold with Two Zinc-Binding Motifs *J. Virol.* 80, 7894-7901.
- Joyce, C.M., and Steitz, T.A. (1995). Polymerase structures and function: variations on a theme? *J Bacteriol* 177, 6321-6329.
- Juuti, J.T., and Bamford, D.H. (1995). RNA binding, packaging and polymerase activities of the different incomplete polymerase complex particles of dsRNA bacteriophage phi 6. *J Mol Biol* 249, 545-554.
- Juuti, J.T., and Bamford, D.H. (1997). Protein P7 of phage phi6 RNA polymerase complex, acquiring of RNA packaging activity by in vitro assembly of the purified protein onto deficient particles. *J Mol Biol* 266, 891-900.

- Juuti, J.T., Bamford, D.H., Tuma, R., and Thomas, G.J., Jr. (1998). Structure and NTPase activity of the RNA-translocating protein (P4) of bacteriophage phi 6. *J Mol Biol* 279, 347-359.
- Kabsch, W. (1993). Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *Journal of Applied Crystallography* 26, 795-800.
- Kao, C.C., Del Vecchio, A.M., and Zhong, W. (1999). De novo initiation of RNA synthesis by a recombinant flaviviridae RNA-dependent RNA polymerase. *Virology* 253, 1-7.
- Kim, M.J., Zhong, W., Hong, Z., and Kao, C.C. (2000). Template nucleotide moieties required for de novo initiation of RNA synthesis by a recombinant viral RNA-dependent RNA polymerase. *J Virol* 74, 10312-10322.
- Kohlstaedt, L.A., Wang, J., Friedman, J.M., Rice, P.A., and Steitz, T.A. (1992). Crystal structure at 3.5 Å resolution of HIV-1 reverse transcriptase complexed with an inhibitor. *Science* 256, 1783-1790.
- Koonin, E.V. (1991). The phylogeny of RNA-dependent RNA polymerases of positive-strand RNA viruses. *J Gen Virol* 72 (Pt 9), 2197-2206.
- Ktistakis, N.T., and Lang, D. (1987). The dodecahedral framework of the bacteriophage phi 6 nucleocapsid is composed of protein P1. *J Virol* 61, 2621-2623.
- Lai, M.M. (1998). Cellular factors in the transcription and replication of viral RNA genomes: a parallel to DNA-dependent RNA transcription. *Virology* 244, 1-12.
- Lampio, A., Kilpelainen, I., Pesonen, S., Karhi, K., Auvinen, P., Somerharju, P., and Kaariainen, L. (2000). Membrane binding mechanism of an RNA virus-capping enzyme. *J Biol Chem* 275, 37853-37859.
- Laurila, M.R., Makeyev, E.V., and Bamford, D.H. (2002). Bacteriophage phi 6 RNA-dependent RNA polymerase: molecular details of initiating nucleic acid synthesis without primer. *J Biol Chem* 277, 17117-17124.
- Laurila, M.R., Salgado, P.S., Stuart, D.I., Grimes, J.M., and Bamford, D.H. (2005). Back-priming mode of phi6 RNA-dependent RNA polymerase. *J Gen Virol* 86, 521-526.
- Lesburg, C.A., Cable, M.B., Ferrari, E., Hong, Z., Mannarino, A.F., and Weber, P.C. (1999). Crystal structure of the RNA-dependent RNA polymerase from hepatitis C virus reveals a fully encircled active site. *Nat Struct Biol* 6, 937-943.
- Love, R.A., Maegley, K.A., Yu, X., Ferre, R.A., Lingardo, L.K., Diehl, W., Parge, H.E., Dragovich, P.S., and Fuhrman, S.A. (2004). The crystal structure of the RNA-dependent RNA polymerase from human rhinovirus: a dual function target for common cold antiviral therapy. *Structure* 12, 1533-1544.

- Luo, G., Hamatake, R.K., Mathis, D.M., Racela, J., Rigat, K.L., Lemm, J., and Colonno, R.J. (2000). De novo initiation of RNA synthesis by the RNA-dependent RNA polymerase (NS5B) of hepatitis C virus. *J Virol* 74, 851-863.
- Makeyev, E.V., and Bamford, D.H. (2000a). The polymerase subunit of a dsRNA virus plays a central role in the regulation of viral RNA metabolism. *EMBO J* 19, 6275-6284.
- Makeyev, E.V., and Bamford, D.H. (2000b). Replicase activity of purified recombinant protein P2 of double-stranded RNA bacteriophage phi6. *EMBO J* 19, 124-133.
- Makeyev, E.V. (2001). PhD Thesis: RNA-Dependent RNA Polymerase of Bacteriophage Φ6. (University of Helsinki, Finland).
- Makeyev, E.V., and Bamford, D.H. (2001). Primer-independent RNA sequencing with bacteriophage phi6 RNA polymerase and chain terminators. *RNA* 7, 774-781.
- Makeyev, E.V., and Grimes, J.M. (2004). RNA-dependent RNA polymerases of dsRNA bacteriophages. *Virus Res* 101, 45-55.
- Malet, H., Egloff, M.P., Selisko, B., Butcher, R.E., Wright, P.J., Roberts, M., Gruez, A., Sulzenbacher, G., Vornrhein, C., Bricogne, G., *et al.* (2007). Crystal structure of the RNA polymerase domain of the West Nile virus non-structural protein 5. *J Biol Chem* 282, 10678-10689.
- Mancini, E.J., Kainov, D.E., Grimes, J.M., Tuma, R., Bamford, D.H., and Stuart, D.I. (2004). Atomic snapshots of an RNA packaging motor reveal conformational changes linking ATP hydrolysis to RNA translocation. *Cell* 118, 743-755.
- Masters, B.S., Stohl, L.L., and Clayton, D.A. (1987). Yeast mitochondrial RNA polymerase is homologous to those encoded by bacteriophages T3 and T7. *Cell* 51, 89-99.
- McCoy, A.J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L.C., Read, R.J. (2007). Phaser crystallographic software. *Journal of Applied Crystallography* 40, 658-674.
- McGraw, T., Mindich, L., and Frangione, B. (1986). Nucleotide sequence of the small double-stranded RNA segment of bacteriophage phi 6: novel mechanism of natural translational control. *J Virol* 58, 142-151.
- Meier, C., Aricescu, A.R., Assenberg, R., Aplin, R.T., Gilbert, R.J., Grimes, J.M., and Stuart, D.I. (2006). The crystal structure of ORF-9b, a lipid binding protein from the SARS coronavirus. *Structure* 14, 1157-1165.
- Mertens, P.P., Diprose, J., Maan, S., Singh, K.P., Attoui, H., and Samuel, A.R. (2004). Bluetongue virus replication, molecular and structural biology. *Vet Ital* 40, 426-437.
- Mindich, L. (1988). Bacteriophage phi 6: a unique virus having a lipid-containing membrane and a genome composed of three dsRNA segments. *Adv Virus Res* 35, 137-176.

- Moll, R., and Hilgenfeld, R. (2004). Bioinformatic approaches suggest several putative functional roles of SARS coronavirus RNA-dependent RNA polymerase. In International Conference on SARS (Lubeck).
- Nelson, C.A., Pekosz, A., Lee, C.A., Diamond, M.S., and Fremont, D.H. (2005). Structure and intracellular targeting of the SARS-coronavirus Orf7a accessory protein. *Structure* 13, 75-85.
- Nettleship, J.E., Brown, J., Groves, M.R., and Geerlof, A. (2008). Methods for protein characterization by mass spectrometry, thermal shift (ThermoFluor) assay, and multiangle or static light scattering. *Methods Mol Biol* 426, 299-318.
- Neufeld, K.L., Galarza, J.M., Richards, O.C., Summers, D.F., and Ehrenfeld, E. (1994). Identification of terminal adenylyl transferase activity of the poliovirus polymerase 3Dpol. *J Virol* 68, 5811-5818.
- Ng, K.K., Cherney, M.M., Vazquez, A.L., Machin, A., Alonso, J.M., Parra, F., and James, M.N. (2002). Crystal structures of active and inactive conformations of a caliciviral RNA-dependent RNA polymerase. *J Biol Chem* 277, 1381-1387.
- Ng, K.K., Pendas-Franco, N., Rojo, J., Boga, J.A., Machin, A., Alonso, J.M., and Parra, F. (2004). Crystal structure of norwalk virus polymerase reveals the carboxyl terminus in the active site cleft. *J Biol Chem* 279, 16638-16645.
- O'Farrell, D., Trowbridge, R., Rowlands, D., and Jager, J. (2003). Substrate complexes of hepatitis C virus RNA polymerase (HC-J4): structural evidence for nucleotide import and de-novo initiation. *J Mol Biol* 326, 1025-1035.
- O'Reilly, E.K., and Kao, C.C. (1998). Analysis of RNA-dependent RNA polymerase structure and function as guided by known polymerase structures and computer predictions of secondary structure. *Virology* 252, 287-303.
- Ojala, P.M., and Bamford, D.H. (1995). In vitro transcription of the double-stranded RNA bacteriophage phi 6 is influenced by purine NTPs and calcium. *Virology* 207, 400-408.
- Ollis, D.L., Brick, P., Hamlin, R., Xuong, N.G., and Steitz, T.A. (1985). Structure of large fragment of Escherichia coli DNA polymerase I complexed with dTMP. *Nature* 313, 762-766.
- Ostermeier, M., Lutz, S., and Benkovic, S.J. (2002). Generation of Protein Fragment Libraries by Incremental Truncation. In Protein-Protein Interactions, E.Golemis, ed. (Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press), pp. 507-522.
- Otwinowski, Z. (1997). Processing of X-ray diffraction data collected in the oscillation mode. *Methods in enzymology* 276, 307-326.
- Palmenberg, A., and Kaesberg, P. (1974). Synthesis of complementary strands of heterologous RNAs with Qbeta replicase. *Proc Natl Acad Sci U S A* 71, 1371-1375.
- Pan, J., Vakharia, V.N., and Tao, Y.J. (2007). The structure of a birnavirus polymerase reveals a distinct active site topology. *Proc Natl Acad Sci U S A* 104, 7385-7390.

- Patton, J.T., and Spencer, E. (2000). Genome replication and packaging of segmented double-stranded RNA viruses. *Virology* 277, 217-225.
- Pedelacq, J.D., Piltch, E., Liong, E.C., Berendzen, J., Kim, C.Y., Rho, B.S., Park, M.S., Terwilliger, T.C., and Waldo, G.S. (2002). Engineering soluble proteins for structural genomics. *Nature Biotechnology* 20, 927-932.
- Pelletier, H., Sawaya, M.R., Kumar, A., Wilson, S.H., and Kraut, J. (1994). Structures of ternary complexes of rat DNA polymerase beta, a DNA template-primer, and ddCTP. *Science* 264, 1891-1903.
- Poch, O., Sauvaget, I., Delarue, M., and Tordo, N. (1989). Identification of four conserved motifs among the RNA-dependent polymerase encoding elements. *EMBO J* 8, 3867-3874.
- Poranen, M.M., Paatero, A.O., Tuma, R., and Bamford, D.H. (2001). Self-assembly of a viral molecular machine from purified protein and RNA constituents. *Mol Cell* 7, 845-854.
- Poranen, M.M., and Tuma, R. (2004). Self-assembly of double-stranded RNA bacteriophages. *Virus Res* 101, 93-100.
- Poranen, M.M., Koivunen, M.R., and Bamford, D.H. (2008a). Nontemplated terminal nucleotidyltransferase activity of double-stranded RNA bacteriophage phi6 RNA-dependent RNA polymerase. *J Virol* 82, 9254-9264.
- Poranen, M.M., Salgado, P.S., Koivunen, M.R., **Wright, S.**, Bamford, D.H., Stuart, D.I., and Grimes, J.M. (2008b). Structural explanation for the role of Mn^{2+} in the activity of phi6 RNA-dependent RNA polymerase. *Nucleic Acids Res* 36, 6633-6644.
- Prentice, E., McAuliffe, J., Lu, X., Subbarao, K., and Denison, M.R. (2004). Identification and characterization of severe acute respiratory syndrome coronavirus replicase proteins. *J Virol* 78, 9977-9986.
- Prodromou, C., Savva, R., and Driscoll, P.C. (2007). DNA fragmentation-based combinatorial approaches to soluble protein expression Part I. Generating DNA fragment libraries. *Drug Discov Today* 12, 931-938.
- Qiao, X., Qiao, J., and Mindich, L. (1997). Stoichiometric packaging of the three genomic segments of double-stranded RNA bacteriophage phi6. *Proc Natl Acad Sci U S A* 94, 4074-4079.
- Ranjith-Kumar, C.T., Gajewski, J., Gutshall, L., Maley, D., Sarisky, R.T., and Kao, C.C. (2001). Terminal nucleotidyl transferase activity of recombinant Flaviviridae RNA-dependent RNA polymerases: implication for viral RNA synthesis. *J Virol* 75, 8615-8623.
- Ranjith-Kumar, C.T., Gutshall, L., Kim, M.J., Sarisky, R.T., and Kao, C.C. (2002a). Requirements for de novo initiation of RNA synthesis by recombinant flaviviral RNA-dependent RNA polymerases. *J Virol* 76, 12526-12536.

- Ranjith-Kumar, C.T., Kim, Y.C., Gutshall, L., Silverman, C., Khandekar, S., Sarisky, R.T., and Kao, C.C. (2002b). Mechanism of de novo initiation by the hepatitis C virus RNA-dependent RNA polymerase: role of divalent metals. *J Virol* 76, 12513-12525.
- Ranjith-Kumar, C.T., Santos, J.L., Gutshall, L.L., Johnston, V.K., Lin-Goerke, J., Kim, M.J., Porter, D.J., Maley, D., Greenwood, C., Earnshaw, D.L., *et al.* (2003). Enzymatic activities of the GB virus-B RNA-dependent RNA polymerase. *Virology* 312, 270-280.
- Ranjith-Kumar, C.T., and Kao, C.C. (2006). Recombinant viral RdRps can initiate RNA synthesis from circular templates. *RNA* 12, 303-312.
- Reinisch, K.M., Nibert, M.L., and Harrison, S.C. (2000). Structure of the reovirus core at 3.6 Å resolution. *Nature* 404, 960-967.
- Ricagno, S., Egloff, M.P., Ulferts, R., Coutard, B., Nurizzo, D., Campanacci, V., Cambillau, C., Ziebuhr, J., and Canard, B. (2006). Crystal structure and mechanistic determinants of SARS coronavirus nonstructural protein 15 define an endoribonuclease family. *Proc. Natl. Acad. Sci. USA* 103, 11892–11897.
- Rohayem, J., Jager, K., Robel, I., Scheffler, U., Temme, A., and Rudolph, W. (2006). Characterization of norovirus 3Dpol RNA-dependent RNA polymerase activity and initiation of RNA synthesis. *J Gen Virol* 87, 2621-2630.
- Rozen, S., and Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* 132, 365-386.
- Saikatendu, K.S., Joseph, J.S., Subramanian, V., Clayton, T., Griffith, M., Moy, K., Velasquez, J., Neuman, B.W., Buchmeier, M.J., Stevens, R.C., and Kuhn, P. (2005). Structural basis of severe acute respiratory syndrome coronavirus ADP-ribose-1"-phosphate dephosphorylation by a conserved domain of nsP3. *Structure* 13, 1665-1675.
- Salgado, P.S., Makeyev, E.V., Butcher, S.J., Bamford, D.H., Stuart, D.I., and Grimes, J.M. (2004). The structural basis for RNA specificity and Ca²⁺ inhibition of an RNA-dependent RNA polymerase. *Structure* 12, 307-316.
- Salgado, P.S. (2005). PhD Thesis: Structural Studies of RdRps. (Oxford University).
- Sambrook, J., Fritsch, E.F., and Maniatis, T. (1989). *Molecular Cloning: A Laboratory Manual* (New York: Cold Spring Harbor Laboratory Press).
- Sarin, L.P., Poranen, M.M., Lehti, N.M., Ravantti, J.J., Koivunen, M.R., Aalto, A.P., van Dijk, A.A., Stuart, D.I., Grimes, J.M., and Bamford, D.H. (2009). Insights into the pre-initiation events of bacteriophage phi 6 RNA-dependent RNA polymerase: towards the assembly of a productive binary complex. *Nucleic Acids Res* 37, 1182-1192.
- Sarin, L.P., **Wright, S.**, Chen, Q., Stuart, D.I., Grimes, J.M., Bamford, D.H. and Poranen, M.M. (2010). The C-terminal domain of bacteriophage φ6 RNA-dependent RNA polymerase regulates transition to elongation. To be submitted.

- Saris, N.E., Mervaala, E., Karppanen, H., Khawaja, J.A., and Lewenstam, A. (2000). Magnesium. An update on physiological, clinical and analytical aspects. *Clin Chim Acta* 294, 1-26.
- Savva, R., Prodromou, C., and Driscoll, P.C. (2007). DNA fragmentation based combinatorial approaches to soluble protein expression Part II: library expression, screening and scale-up. *Drug Discov Today* 12, 939-947.
- Sawaya, M.R., Pelletier, H., Kumar, A., Wilson, S.H., and Kraut, J. (1994). Crystal structure of rat DNA polymerase beta: evidence for a common polymerase mechanism. *Science* 264, 1930-1935.
- Schatz, P. (1993). Use of peptide libraries to map the substrate specificity of a peptide-modifying enzyme: a 13 residue consensus peptide specifies biotinylation in *Escherichia coli*. *Biotechnology* 11, 1138-1143.
- Shim, J.H., Larson, G., Wu, J.Z., and Hong, Z. (2002). Selection of 3'-template bases and initiating nucleotides by hepatitis C virus NS5B RNA-dependent RNA polymerase. *J Virol* 76, 7030-7039.
- Snijder EJ, Bredenbeek PJ, Dobbe JC, Thiel V, Ziebuhr J, Poon LL, Guan Y, Rozanov M, Spaan WJ, and Gorbalenya AE (2003). Unique and conserved features of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 lineage. *Journal of Molecular Biology* 331, 991-1004.
- Sousa, R., Chung, Y.J., Rose, J.P., and Wang, B.C. (1993). Crystal structure of bacteriophage T7 RNA polymerase at 3.3 Å resolution. *Nature* 364, 593-599.
- Stadler K, Massignani V, Eickmann M, Becker S, Abrignani S, Klenk HD, and Rappuoli R (2003). SARS - beginning to understand a new virus. *Nature Reviews Microbiology* 1, 209-218.
- Starnes, M.C., and Joklik, W.K. (1993). Reovirus protein lambda 3 is a poly(C)-dependent poly(G) polymerase. *Virology* 193, 356-366.
- Steitz, T.A., and Steitz, J.A. (1993). A general two-metal-ion mechanism for catalytic RNA. *Proc Natl Acad Sci U S A* 90, 6498-6502.
- Steitz, T.A. (1998). A mechanism for all polymerases. *Nature* 391, 231-232.
- Su, D., Lou, Z., Sun, F., Zhai, Y., Yang, H., Zhang, R., Joachimiak, A., Zhang, X.C., Bartlam, M., and Rao, Z. (2006). Dodecamer Structure of Severe Acute Respiratory Syndrome Coronavirus Nonstructural Protein nsp10. *J. Virol.* 80, 7902-7908.
- Sun, J.H., Adkins, S., Faurete, G., and Kao, C.C. (1996). Initiation of (-)-strand RNA synthesis catalyzed by the BMV RNA-dependent RNA polymerase: synthesis of oligonucleotides. *Virology* 226, 1-12.

- Sun, X.L., Johnson, R.B., Hockman, M.A., and Wang, Q.M. (2000). De novo RNA synthesis catalyzed by HCV RNA-dependent RNA polymerase. *Biochem Biophys Res Commun* 268, 798-803.
- Sutton, G., Fry, E., Carter, L., Sainsbury, S., Walter, T., Nettleship, J., Berrow, N., Owens, R., Gilbert, R., Davidson, A., *et al.* (2004). The nsp9 replicase protein of SARS-coronavirus, structure and functional insights. *Structure* 12, 341-353.
- Tabor, S., and Richardson, C.C. (1989). Effect of manganese ions on the incorporation of dideoxynucleotides by bacteriophage T7 DNA polymerase and Escherichia coli DNA polymerase I. *Proc Natl Acad Sci U S A* 86, 4076-4080.
- Tao, Y., Farsetta, D.L., Nibert, M.L., and Harrison, S.C. (2002). RNA synthesis in a cage--structural studies of reovirus polymerase lambda3. *Cell* 111, 733-745.
- Tarendeau, F., Boudet, J., Guilligay, D., Mas, P.J., Bougault, C.M., Boulo, S., Baudin, F., Ruigrok, R.W., Daigle, N., Ellenberg, J., *et al.* (2007). Structure and nuclear import function of the C-terminal domain of influenza virus polymerase PB2 subunit. *Nat Struct Mol Biol* 14, 229-233.
- Thompson, A.A., and Peersen, O.B. (2004). Structural basis for proteolysis-dependent activation of the poliovirus RNA-dependent RNA polymerase. *EMBO J* 23, 3462-3471.
- Thompson, A.A., Albertini, R.A., and Peersen, O.B. (2007). Stabilization of poliovirus polymerase by NTP binding and fingers-thumb interactions. *J Mol Biol* 366, 1459-1474.
- Torralba, S., Sojat, J., and Hartmann, R. (2008). 2'-5' oligoadenylate synthetase shares active site architecture with the archaeal CCA-adding enzyme. *Cell Mol Life Sci* 65, 2613-2620.
- van Dijk, A.A., Frilander, M., and Bamford, D.H. (1995). Differentiation between minus- and plus-strand synthesis: polymerase activity of dsRNA bacteriophage phi 6 in an in vitro packaging and replication system. *Virology* 211, 320-323.
- van Dijk, A.A., and Huismans, H. (1980). The in vitro activation and further characterization of the bluetongue virus-associated transcriptase. *Virology* 104, 347-356.
- van Dijk, A.A., Makeyev, E.V., and Bamford, D.H. (2004). Initiation of viral RNA-dependent RNA polymerization. *J Gen Virol* 85, 1077-1093.
- Vassilyev, D.G., Sekine, S., Laptenko, O., Lee, J., Vassilyeva, M.N., Borukhov, S., and Yokoyama, S. (2002). Crystal structure of a bacterial RNA polymerase holoenzyme at 2.6 Å resolution. *Nature* 417, 712-719.
- Vazquez, A.L., Martin Alonso, J.M., Casais, R., Boga, J.A., and Parra, F. (1998). Expression of enzymatically active rabbit hemorrhagic disease virus RNA-dependent RNA polymerase in Escherichia coli. *J Virol* 72, 2999-3004.

- von Brunn, A., Teepe, C., Simpson, J.C., Pepperkok, R., Friedel, C.C., Zimmer, R., Roberts, R., Baric, R., and Haas, J. (2007). Analysis of intraviral protein-protein interactions of the SARS coronavirus ORFeome. *PLoS ONE* 2, e459.
- Waldo, G.S., Standish, B.M., Berendzen, J., and Terwilliger, T.C. (1999). Rapid protein-folding assay using green fluorescent protein. *Nature Biotechnology* 17, 691-695.
- Waldo, G.S. (2003). Genetic screens and directed evolution for protein solubility. *Curr Opin Chem Biol* 7, 33-38.
- Walter, T.S., Diprose, J.M., Mayo, C.J., Siebold, C., Pickford, M.G., Carter, L., Sutton, G.C., Berrow, N.S., Brown, J., Berry, I.M., *et al.* (2005). A procedure for setting up high-throughput nanolitre crystallization experiments. Crystallization workflow for initial screening, automated storage, imaging and optimization. *Acta Crystallogr D Biol Crystallogr* 61, 651-657.
- Wang, T.S., and Korn, D. (1982). Specificity of the catalytic interaction of human DNA polymerase beta with nucleic acid substrates. *Biochemistry* 21, 1597-1608.
- Westover, K.D., Bushnell, D.A., and Kornberg, R.D. (2004). Structural basis of transcription: separation of RNA from DNA by RNA polymerase II. *Science* 303, 1014-1016.
- Wright, S.**, Poranen, M.M., Bamford, D.H., Stuart, D.I., Grimes, J.M. (2010). Non-catalytic ions direct viral RNA-dependent RNA polymerase from de novo initiation to elongation. *PLoS One*. Submitted.
- Xiong, Y., and Eickbush, T.H. (1990). Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J* 9, 3353-3362.
- Xu, X., Liu, Y., Weiss, S., Arnold, E., Sarafianos, S.G., and Ding, J. (2003). Molecular model of SARS coronavirus polymerase: implications for biochemical functions and drug design. *Nucleic Acids Res* 31, 7117-7130.
- Xu, X., Zhai, Y., Sun, F., Lou, Z., Su, D., Xu, Y., Zhang, R., Joachimiak, A., Zhang, X.C., Bartlam, M., and Rao, Z. (2006). New antiviral target revealed by the hexameric structure of mouse hepatitis virus nonstructural protein nsp15. *J Virol* 80, 7909-7917..
- Yang, H., Makeyev, E.V., and Bamford, D.H. (2001). Comparison of polymerase subunits from double-stranded RNA bacteriophages. *J Virol* 75, 11088-11095.
- Yang H, Yang M, Ding Y, Liu Y, Lou Z, Zhou Z, Sun L, Mo L, Ye S, Pang H, *et al.* (2003). The crystal structures of severe acute respiratory syndrome virus main protease and its complex with an inhibitor. *Proceedings of the National Academy of Sciences of the United States of America* 100, 13190-13195.
- Yang, H., Makeyev, E.V., Butcher, S.J., Gaidelyte, A., and Bamford, D.H. (2003). Two distinct mechanisms ensure transcriptional polarity in double-stranded RNA bacteriophages. *J Virol* 77, 1195-1203.

- Yap, T.L., Xu, T., Chen, Y.L., Malet, H., Egloff, M.P., Canard, B., Vasudevan, S.G., and Lescar, J. (2007). Crystal structure of the dengue virus RNA-dependent RNA polymerase catalytic domain at 1.85-angstrom resolution. *J Virol* 81, 4753-4765.
- Yi, G.H., Zhang, C.Y., Cao, S., Wu, H.X., and Wang, Y. (2003). De novo RNA synthesis by a recombinant classical swine fever virus RNA-dependent RNA polymerase. *Eur J Biochem* 270, 4952-4961.
- Zamyatkin, D.F., Parra, F., Alonso, J.M., Harki, D.A., Peterson, B.R., Grochulski, P., and Ng, K.K. (2008). Structural insights into mechanisms of catalysis and inhibition in Norwalk virus polymerase. *J Biol Chem* 283, 7705-7712.
- Zamyatkin, D.F., Parra, F., Machin, A., Grochulski, P., and Ng, K.K. (2009). Binding of 2'-amino-2'-deoxycytidine-5'-triphosphate to norovirus polymerase induces rearrangement of the active site. *J Mol Biol* 390, 10-16.
- Zhai, Y., Sun, F., Li, X., Pang, H., Xu, X., Bartlam, M., and Rao, Z. (2005). Insights into SARS-CoV transcription and replication from the structure of the nsp7-nsp8 hexadecamer. *Nat. Struct. Mol. Biol.* 12, 980-986.
- Zhang, G., Campbell, E.A., Minakhin, L., Richter, C., Severinov, K., and Darst, S.A. (1999). Crystal structure of *Thermus aquaticus* core RNA polymerase at 3.3 Å resolution. *Cell* 98, 811-824.
- Zhang, R.Q., and Ellis, K.J. (1989). In vivo measurement of total body magnesium and manganese in rats. *Am J Physiol* 257, R1136-1140.
- Zhong, W., Gutshall, L.L., and Del Vecchio, A.M. (1998). Identification and characterization of an RNA-dependent RNA polymerase activity within the nonstructural protein 5B region of bovine viral diarrhea virus. *J Virol* 72, 9365-9369.
- Zhong, W., Uss, A.S., Ferrari, E., Lau, J.Y., and Hong, Z. (2000). De novo initiation of RNA synthesis by hepatitis C virus nonstructural protein 5B polymerase. *J Virol* 74, 2017-2022.
- Ziebuhr, J. (2005). The coronavirus replicase. *Curr. Top. Microbiol. Immunol.* 287, 57-94.

Appendix

A1.1. pOPIN Vectors

The Oxford Protein Production Facility (OPPF) has developed a suite of In-fusion vectors to allow the expression of proteins with a range of different purification/solubility tags in different heterologous hosts (*E. coli*, insect and mammalian cells). Specific details regarding pOPINF, pOPINJ and pOPINS vectors used in this ESPRIT work are given in **Table A1.1**.

Vectors pOPINF (addition of N-His₆ tag) and pOPINJ (addition of N-His₆-GST tag) are based on the pTriEx2 vector. They are T7/*lacO* driven expression vectors for use with (DE3) *E. coli* strains, but can also exploit their hybrid cytomegalovirus (CMV) enhancer and β -actin promoter to drive mammalian expression, or use the baculoviral p10 promoter and the *lef-2* and 1629 baculo-recombination elements for insect cell expression; an overview of plasmid elements for pOPINF are shown in **Fig. A1.1**. pOPINF and pOPINJ confer Ampicillin/Carbenicillin resistance and are considered high copy number plasmids. Incubation with human rhinovirus (HRV) 3C protease is required for tag removal.

The pOPINS vector (addition of N-His₆-SUMO tag) is based on the pET28a vector. These are T7/*lacO* driven expression vectors and must be used in a (DE3) *E. coli* strain. pOPINS confers Kanamycin resistance and is considered a low copy number plasmid. Incubation with SUMO protease is required for tag removal.

Table A1.1. Vectors used in ESPRIT experiments.

Vector	Fusion Tag	Product	Parent Vector/ Antibiotic resistance	Forward Primer Extension	Reverse Primer Extension
pOPINF	MAHHHHHHS SGLEVLFG GP	N ^{His} -3C-POI	pTriEx2/Ampicillin	AAGTTCTGT TTCAGGGCC CG	ATGGTCTA GAAAGCT <u>T</u> <u>TA</u>
pOPINS	... MGSSHHHHH H-SUMO ...	N ^{His} -SUMO- 3C-POI	pET28a/Kanamycin	GCGAACAG ATCGGTGGT	ATGGTCTA GAAAGCT <u>T</u> <u>TA</u>
pOPINJ	MAHHHHHHS SG-GST- LEVLFG GP...	N ^{His} -GST-3C- POI	pTriEx2/Ampicillin	AAGTTCTGT TTCAGGGCC CG	ATGGTCTA GAAAGCT <u>T</u> <u>TA</u>

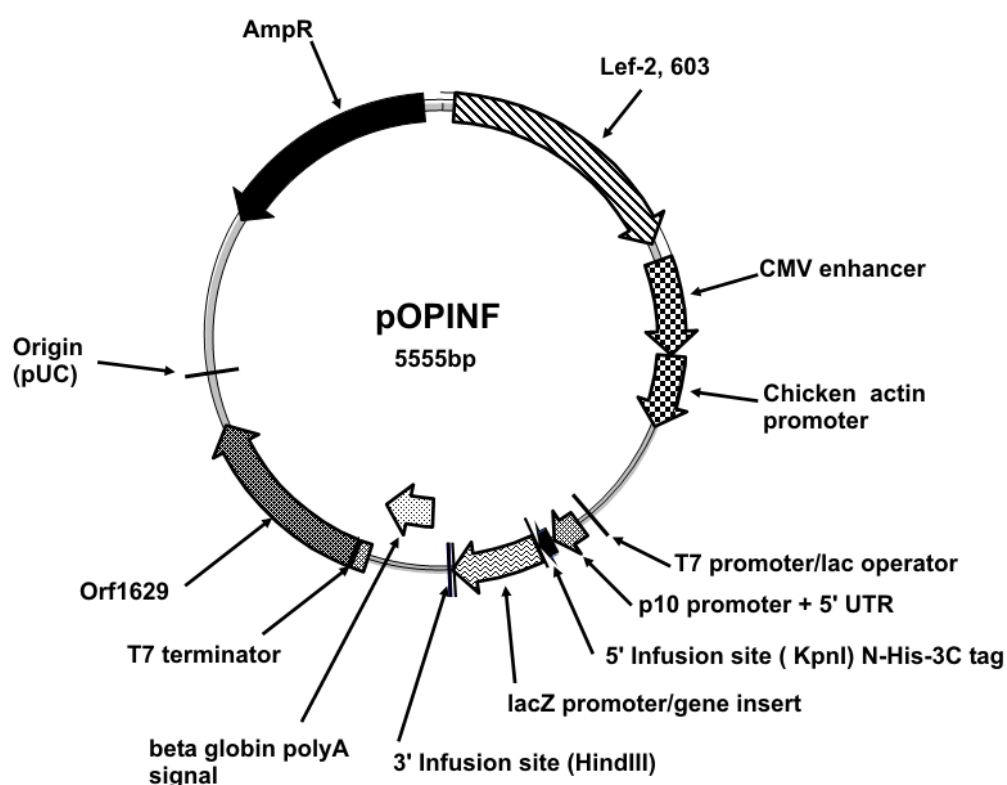
| is the cleavage site for human rhinovirus 3C protease

| is the cleavage site for SUMO protease

Underlined bases mark termination codons

Figure A1.1. Plasmid map of pOPINF.

The pOPIN vectors contain sequences that permit expression in *E. coli*, insect cells and mammalian cells. In addition they contain a *LacZ* gene insert encoding β -galactosidase, which is flanked by the 3' and 5' In-Fusion sites to allow for blue/white selection when bromo-chloro-indolyl-galactopyranoside (X-Gal) and IPTG are present in the LB-Agar media alongside the antibiotic. Any re-ligated vectors (without gene of interest inserted) that are transformed into *E. coli* will express β -galactosidase and convert X-Gal into 5,5'-dibromo-4,4'-dichloro-indigo, an insoluble blue product making colonies stand out as blue.



A1.2. Primer design and In-Fusion

Gene truncation specific primers were designed from the sequencing results and were appended with vector specific extensions as listed in **Table A1.1**. For the synthetic construct (Syn, 433-841), primers were generated semi-automatically from the codon-optimised SARS NSP-12 gene, using the OPPF Opine primer design utility that runs Primer3 (Rozen and Skaletsky, 2000) to design an appropriate target-specific primer annealing region to which the vector-specific extensions shown in **Table A1.1** were added. The PCR reaction, PCR product purification and In-Fusion reaction (Clontech) was then carried out as described in section 4.6.2.1. A schematic of the In-Fusion reaction technology is shown in **Fig. A1.2**.

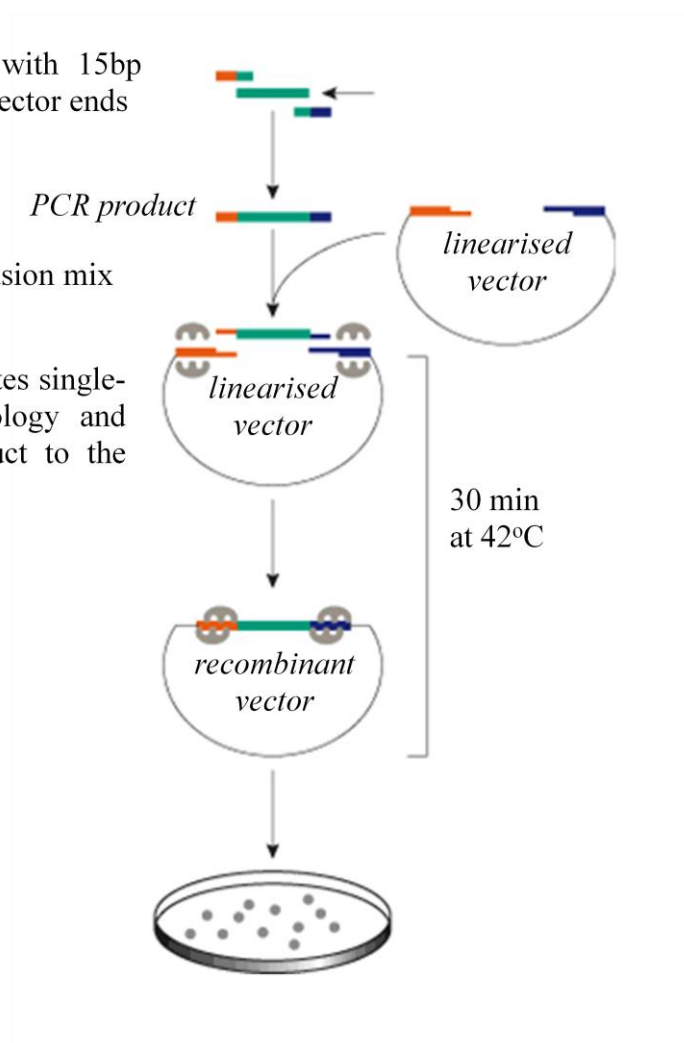
Figure A1.2. Single-step cloning using In-Fusion technology (Clontech).

This method uses a proprietary enzyme, which recognises regions of homology between a PCR product and the ends of a linearised vector. The gene of interest is PCR-amplified with flanking sequences that are identical to the ends of linearised target vector (15bp or greater). Incubation at 42°C for 30 min with In-Fusion enzyme generates regions of single-stranded DNA at the ends of PCR product and vector, which can anneal. After transformation and antibiotic and blue/white selection in *E. coli*, intact expression vector is obtained. Figure adapted from In-Fusion cloning system manual (Clontech).

(i) Gene-specific primers with 15bp extensions homologous to vector ends

(ii) Add to lyophilised In-Fusion mix

(iii) In-Fusion enzyme creates single-stranded regions of homology and then fuses the PCR product to the vector



Publications

Poranen, M.M., Salgado, P.S., Koivunen, M.R., **Wright, S.**, Bamford, D.H., Stuart, D.I., and Grimes, J.M. (2008b). Structural explanation for the role of Mn^{2+} in the activity of phi6 RNA-dependent RNA polymerase. *Nucleic Acids Res* 36, 6633-6644.

Wright, S., Poranen, M.M., Bamford, D.H., Stuart, D.I., Grimes, J.M. (2010). Non-catalytic ions direct viral RNA-dependent RNA polymerase from de novo initiation to elongation. [Submitted].

Sarin, L.P., **Wright, S.**, Chen, Q., Stuart, D.I., Grimes, J.M., Bamford, D.H. and Poranen, M.M. (2010). The C-terminal domain of bacteriophage $\phi 6$ RNA-dependent RNA polymerase regulates transition to elongation. [Submitted].